Mathematical Statistics

*Old School*

John I. Marden
Department of Statistics
University of Illinois at Urbana-Champaign
# Contents

1 Introduction
   1.1 Examples
      1.1.1 Example – Fruit Flies
      1.1.2 Example – Political Interest
      1.1.3 Example – Mouth Size

I Distribution Theory

2 Distributions and Densities
   2.1 Probability models
   2.2 Distribution functions
   2.3 Densities
      2.3.1 Probability density functions
      2.3.2 Probability mass functions
   2.4 Distributions without pdf’s or pmf’s
      2.4.1 Example: Late start
      2.4.2 Example: Spinner
      2.4.3 Example: Random coin

3 Expected Values
   3.1 Definition
   3.2 Means, Variances, and Covariances
      3.2.1 Example
      3.2.2 The variance of linear combinations & affine transformations
   3.3 Other summaries
   3.4 The moment and cumulant generating functions
      3.4.1 Example: Normal
      3.4.2 Example: Gamma
      3.4.3 Example: Binomial and Multinomial
   3.5 Vectors and matrices

iii
4 Marginal Distributions and Independence 35
4.1 Marginal distributions ........................................... 35
4.1.1 Example: Marginals of a multinomial .................... 36
4.2 Marginal densities ............................................. 36
4.2.1 Example: Ranks ............................................ 37
4.3 Independence ................................................ 39
4.3.1 Example: Independent exponentials ...................... 40
4.3.2 Spaces and densities ..................................... 40
4.3.3 IID ......................................................... 43

5 Transformations I 45
5.1 Distribution functions ............................................ 45
5.1.1 Example: Sum of uniforms ................................ 45
5.1.2 Example: Chi-squared on 1 df ............................ 47
5.1.3 Example: Probability transform ......................... 48
5.2 Moment generating functions ................................. 49
5.2.1 Example: Uniform → Exponential ...................... 50
5.2.2 Example: Sums of independent Gamma’s ............... 50
5.2.3 Linear combinations of independent Normals .......... 50
5.2.4 Normalized means ....................................... 51
5.3 Discrete distributions .......................................... 53
5.3.1 Example: Sum of discrete uniforms ..................... 53
5.3.2 Example: Convolutions ................................. 54
5.3.3 Example: Sum of two Poissons ......................... 55
5.3.4 Binomials and Bernoullis ............................... 56

6 Transformations II: Continuous distributions 59
6.1 One dimension .................................................. 59
6.2 General case .................................................. 60
6.2.1 Example: Gamma – Beta ................................ 61
6.2.2 Example: The Dirichlet distribution .................... 63
6.2.3 Example: Affine transformations ....................... 65
6.2.4 Example: Bivariate Normal .............................. 65
6.2.5 Example: Orthogonal transformations .................. 66
6.2.6 Example: Polar coordinates .............................. 67
6.3 Order statistics ................................................ 69
6.3.1 Approximating the mean and variance ................. 71

7 Conditional distributions 73
7.1 Introduction ................................................... 73
7.2 Examples of conditional distributions ....................... 74
7.2.1 Simple linear model .................................... 74
7.2.2 Finite mixture model .................................... 75
7.2.3 Hierarchical models ..................................... 75
7.2.4 Bayesian models ....................................... 75
7.3 Conditional + Marginal → Joint .............................. 76
7.3.1 Joint densities ........................................... 77
7.4 The marginal distribution of Y ............................... 78
7.4.1 Example: Coins and the Beta-Binomial ............... 78
7.4.2 Example: Simple normal linear model ................. 78
7.4.3 Marginal mean and variance . . . . . . . . . . . . . . . . 79
7.4.4 Example: Fruit flies, continued . . . . . . . . . . . . . . . . 81
7.5 The conditional from the joint . . . . . . . . . . . . . . . . . . . . 82
7.5.1 Example: Coin . . . . . . . . . . . . . . . . . . . . . . . . . . . . 82
7.5.2 Bivariate normal . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83
7.6 Bayes Theorem: Reversing the conditionals . . . . . . . . . . . . 85
7.6.1 Example: AIDS virus . . . . . . . . . . . . . . . . . . . . . . . . . 86
7.6.2 Example: Beta posterior for the binomial . . . . . . . . . . . 86
7.7 Conditionals and independence . . . . . . . . . . . . . . . . . . . . . 86
7.7.1 Example: Independence of residuals and $X$ . . . . . . . . . . 87

8 The Multivariate Normal Distribution 89
8.1 Definition . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
8.1.1 The spectral decomposition . . . . . . . . . . . . . . . . . . . . . . 91
8.2 Some properties of the multivariate normal . . . . . . . . . . . . 92
8.2.1 Affine transformations . . . . . . . . . . . . . . . . . . . . . . . . . 92
8.2.2 Marginals . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92
8.2.3 Independence . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92
8.3 The pdf . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 93
8.4 The sample mean and variance . . . . . . . . . . . . . . . . . . . . . . 94
8.5 Chi-squares . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 96
8.5.1 Moments of a chi-squared . . . . . . . . . . . . . . . . . . . . . . . 98
8.5.2 Other covariance matrices . . . . . . . . . . . . . . . . . . . . . . . 98
8.6 Student’s $t$ distribution . . . . . . . . . . . . . . . . . . . . . . . . . . 100
8.7 Linear models and the conditional distribution . . . . . . . . . . . . 101

II Statistical Inference 103

9 Statistical Models and Inference 105
9.1 Statistical models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 105
9.2 Interpreting probability . . . . . . . . . . . . . . . . . . . . . . . . . . . 106
9.3 Approaches to inference . . . . . . . . . . . . . . . . . . . . . . . . . . . 108

10 Introduction to Estimation 111
10.1 Definition of estimator . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111
10.2 Plug-in methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 112
10.3 Least squares . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113
10.4 The likelihood function . . . . . . . . . . . . . . . . . . . . . . . . . . . 114
10.4.1 Maximum likelihood estimation . . . . . . . . . . . . . . . . . . . 115
10.5 The posterior distribution . . . . . . . . . . . . . . . . . . . . . . . . . 117
10.6 Bias, variance, and MSE . . . . . . . . . . . . . . . . . . . . . . . . . . 118
10.7 Functions of estimators . . . . . . . . . . . . . . . . . . . . . . . . . . . 120
10.7.1 Example: Poisson . . . . . . . . . . . . . . . . . . . . . . . . . . . 120

11 Likelihood and sufficiency 123
11.1 The likelihood principle . . . . . . . . . . . . . . . . . . . . . . . . . . . 123
11.1.1 Binomial and negative binomial . . . . . . . . . . . . . . . . . . . 124
11.2 Sufficiency . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125
11.2.1 Examples . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 126
Contents

11.3 Conditioning on a sufficient statistic ............................................ 128
  11.3.1 Example: IID .......................................................... 131
  11.3.2 Example: Normal mean ................................................. 131
11.4 Rao-Blackwell: Improving an estimator ......................................... 132
  11.4.1 Example: Normal probability .......................................... 133
  11.4.2 Example: IID .......................................................... 134

12 UMVUE's .................................................................................. 135
  12.1 Unbiased estimation .................................................................. 135
  12.2 Completeness: Unique unbiased estimators ....................................... 137
    12.2.1 Example: Poisson ...................................................... 139
    12.2.2 Example: Uniform .................................................... 139
  12.3 Uniformly minimum variance estimators ......................................... 140
    12.3.1 Example: Poisson ...................................................... 141
  12.4 Completeness for exponential families ........................................... 141
    12.4.1 Examples .................................................................. 142
  12.5 Fisher's Information and the Cramér-Rao Lower Bound ....................... 143
    12.5.1 Example: Double exponential ....................................... 144
    12.5.2 Example: Normal $\mu^2$ .............................................. 145

13 Location Families and Shift Equivariance ........................................ 147
  13.1 Introduction ............................................................................ 147
  13.2 Location families .................................................................... 148
  13.3 Shift-equivariant estimators ........................................................ 149
    13.3.1 The Pitman estimator .................................................. 149
  13.4 Examples .............................................................................. 152
    13.4.1 Exponential .............................................................. 152
    13.4.2 Double Exponential ................................................... 153
  13.5 General invariance and equivariance .............................................. 154
    13.5.1 Groups and their actions .............................................. 154
    13.5.2 Invariant models ....................................................... 156
    13.5.3 Induced actions ......................................................... 157
    13.5.4 Equivariant estimators ............................................... 158
    13.5.5 Groups ................................................................. 158

14 The Decision-theoretic Approach ...................................................... 159
  14.1 The setup ............................................................................. 159
  14.2 The risk function .................................................................... 160
  14.3 Admissibility .......................................................................... 160
    14.3.1 Stein's surprising result ............................................... 162
  14.4 Bayes procedures ................................................................... 165
    14.4.1 Bayes procedures and admissibility ................................... 166
  14.5 Minimax procedures ................................................................ 166
    14.5.1 Example: Binomial .................................................... 167

15 Asymptotics: Convergence in probability .......................................... 169
  15.1 The set-up ............................................................................ 169
  15.2 Convergence in probability to a constant ...................................... 170
    15.2.1 Chebyshev's Inequality .............................................. 171
  15.3 The law of large numbers; Mapping ............................................ 173
## 16 Asymptotics: Convergence in distribution

16.1 Converging to a constant random variable ........................................ 179
16.2 Moment generating functions ............................................................ 180
16.3 The Central Limit Theorem ............................................................... 180
16.3.1 Supersizing ....................................................................................... 181
16.4 Mapping ............................................................................................... 183
16.5 The Δ-method ....................................................................................... 185
16.5.1 The median ....................................................................................... 186
16.5.2 Variance stabilizing transformations ............................................... 187
16.6 The MLE in Exponential Families .......................................................... 188
16.7 The multivariate Δ-method ................................................................... 190

## 17 Likelihood Estimation

17.1 The setup: Cramér’s conditions ............................................................. 195
17.2 Consistency .......................................................................................... 197
17.2.1 Convexity and Jensen’s inequality .................................................... 197
17.2.2 A consistent sequence of roots ......................................................... 198
17.3 Asymptotic normality ........................................................................... 200
17.4 Asymptotic efficiency .......................................................................... 203
17.5 Multivariate parameters ...................................................................... 204

## 18 Hypothesis Testing

18.1 Accept/Reject ....................................................................................... 210
18.2 Test functions ..................................................................................... 213
18.3 Simple versus simple .......................................................................... 214
18.4 The Neyman-Pearson Lemma .............................................................. 217
18.4.1 Examples ........................................................................................ 218
18.4.2 Bayesian approach ........................................................................ 222
18.5 Uniformly most powerful tests .............................................................. 223
18.5.1 One-sided exponential family testing problems ................................ 226
18.5.2 Monotone likelihood ratio ............................................................... 227
18.6 Some additional criteria ....................................................................... 227

## 19 Likelihood-based testing

19.1 Score tests .......................................................................................... 230
19.1.1 One-sided ....................................................................................... 230
19.1.2 Many-sided ..................................................................................... 231
19.2 The maximum likelihood ratio test ...................................................... 233
19.2.1 Examples ........................................................................................ 234
19.2.2 Asymptotic null distribution ........................................................... 236

## 20 Nonparametric Tests and Randomization

20.1 Testing the median ............................................................................... 242
20.1.1 A randomization distribution .......................................................... 243
20.2 Randomization test formalities .............................................................. 247
20.2.1 Using p-values ............................................................................... 249
20.2.2 Using the normal approximation ..................................................... 249
20.2.3 When the regular and randomization distributions coincide ............ 250
20.2.4 Example: The signed rank test ....................................................... 250
20.3 Two-sample tests ............................................................................... 252
Formally, a statistical model is a family of probability distributions on a sample space. In practice, the model is supposed to be an abstraction of some real (scientific) process or experiment. The sample space specifies the set of potential outcomes of the experiment (the data), and the presumption is that one of the members of the family of probability distributions describes the chances of the various outcomes. If the distribution were known, then the model would be a probability model. What makes it a statistical model is that the distribution is unknown, and one wishes to infer something about the distribution, or make some kind of decision, based on observing the data. Choosing the model for a given situation is not always easy. It is important to both be very specific about what is being assumed, and to be well aware that the model is wrong. The hope is that the model forms a basis for inference that comes reasonably close to reality.

We will start with some examples. These are small data sets with fairly simple models, but they give an idea of the type of situations we are considering. Section 2.1 formally presents probability models, and Section 9.1 presents statistical models. Subsequent sections cover some general types of models.

1.1 Examples

1.1.1 Example – Fruit Flies

Arnold [1981] presents an experiment concerning the genetics of Drosophila pseudoobscura, a type of fruit fly. We are looking at a particular locus (place) on a pair of chromosomes. The locus has two possible alleles (values): TL ≡ TreeLine and CU ≡ Cuernavaca. Each individual has two of these, one on each chromosome. The individual’s genotype is the pair of alleles it has. Thus the genotype could be (TL,TL), (TL,CU), or (CU,CU). [There is no distinction made between (CU,TL) and (TL,CU).]

The objective is to estimate $\theta \in (0,1)$, the proportion of CU in the population. In this experiment, the researchers randomly collected 10 adult males. Unfortunately, one cannot determine the genotype of the adult fly just by looking at him. One can determine the genotype of young flies, though. So the researchers bred each of these ten flies with a (different) female known to be (TL,TL), and analyzed two of the offspring from each mating. The genotypes of these offspring are given in Table 1.1.
Chapter 1. Introduction

Table 1.1: The data on the fruit flies

<table>
<thead>
<tr>
<th>Father</th>
<th>Offsprings’ Genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(TL,TL) &amp; (TL,TL)</td>
</tr>
<tr>
<td>2</td>
<td>(TL,TL) &amp; (TL,CU)</td>
</tr>
<tr>
<td>3</td>
<td>(TL,TL) &amp; (TL,TL)</td>
</tr>
<tr>
<td>4</td>
<td>(TL,TL) &amp; (TL,TL)</td>
</tr>
<tr>
<td>5</td>
<td>(TL,CU) &amp; (TL,CU)</td>
</tr>
<tr>
<td>6</td>
<td>(TL,TL) &amp; (TL,CU)</td>
</tr>
<tr>
<td>7</td>
<td>(TL,CU) &amp; (TL,CU)</td>
</tr>
<tr>
<td>8</td>
<td>(TL,TL) &amp; (TL,TL)</td>
</tr>
<tr>
<td>9</td>
<td>(TL,CU) &amp; (TL,CU)</td>
</tr>
<tr>
<td>10</td>
<td>(TL,TL) &amp; (TL,TL)</td>
</tr>
</tbody>
</table>

The sample space consists of all possible combinations of 10 sets of two pairs of alleles, each pair being either (TL,TL) or (TL,CU). The probability of these outcomes is governed by the population proportion of CU’s, but some assumptions are necessary to specify the probabilities:

1. The ten chosen Fathers are a simple random sample from the population.
2. The chance that a given Father has 0, 1 or 2 CU’s in his genotype follows the Hardy-Weinberg laws, which means that the number of CU’s is like flipping a coin twice independently, with probability of heads being \( \theta \).
3. For a given mating, the two offspring are each equally likely to get either of the Father’s two alleles (as well as a TL from the Mother), and what the two offspring get are independent.

Based on these assumptions, the goal would be to figure out the probability distribution, and to find an estimate of \( \theta \) and its variability, e.g., a 95% confidence interval. One should also try to assess the plausibility of the assumptions, although with a small data set like this one it may be difficult.

1.1.2 Example – Political Interest

Lazarsfeld, Berelson and Gaudet (1968, *The People’s Choice*) collected some data to determine the relationship between level of education and intention to vote in an election. The variables of interest were

- \( X \) = Education: 0 = Some high school, 1 = No high school
- \( Y \) = Interest: 0 = Great political interest, 1 = Moderate political interest, 2 = No political interest
- \( Z \) = Vote: 0 = Intends to vote, 1 = Does not intend to vote

Here is the table of data:
You would expect $X$ and $Z$ to be dependent, that is, people with more education are more likely to vote. That’s not the question. The question is whether education and voting are conditionally independent given interest, that is, once you know someone’s level of political interest, knowing their educational level does not help you predict whether they vote.

The model is that the $n = 2812$ observations are independent and identically distributed, with each having the same set of probabilities of landing in any one of the twelve possible cells of the table. Based on this model, how can the hypothesis of interest be expressed? How can it be tested, given the data?

### 1.1.3 Example – Mouth Size

Measurements were made on the size of mouths of 27 kids at four ages: 8, 10, 12, and 14\(^1\). The measurement is the distance from the “center of the pituitary to the pterygomaxillary fissure\(^2\)” in millimeters.

There are 11 girls (Sex=1) and 16 boys (Sex=0). Here is the raw data:

<table>
<thead>
<tr>
<th>Age8</th>
<th>Age10</th>
<th>Age12</th>
<th>Age14</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21.0</td>
<td>20.0</td>
<td>21.5</td>
<td>23.0</td>
</tr>
<tr>
<td>2</td>
<td>21.0</td>
<td>21.5</td>
<td>24.0</td>
<td>25.5</td>
</tr>
<tr>
<td>3</td>
<td>20.5</td>
<td>24.0</td>
<td>24.5</td>
<td>26.0</td>
</tr>
<tr>
<td>4</td>
<td>23.5</td>
<td>24.5</td>
<td>25.0</td>
<td>26.5</td>
</tr>
<tr>
<td>5</td>
<td>21.5</td>
<td>23.0</td>
<td>22.5</td>
<td>23.5</td>
</tr>
<tr>
<td>6</td>
<td>20.0</td>
<td>21.0</td>
<td>21.0</td>
<td>22.5</td>
</tr>
<tr>
<td>7</td>
<td>21.5</td>
<td>22.5</td>
<td>23.0</td>
<td>25.0</td>
</tr>
<tr>
<td>8</td>
<td>23.0</td>
<td>23.0</td>
<td>23.5</td>
<td>24.0</td>
</tr>
<tr>
<td>9</td>
<td>20.0</td>
<td>21.0</td>
<td>22.0</td>
<td>21.5</td>
</tr>
<tr>
<td>10</td>
<td>16.5</td>
<td>19.0</td>
<td>19.0</td>
<td>19.5</td>
</tr>
<tr>
<td>11</td>
<td>24.5</td>
<td>25.0</td>
<td>28.0</td>
<td>28.0</td>
</tr>
<tr>
<td>12</td>
<td>26.0</td>
<td>25.0</td>
<td>29.0</td>
<td>31.0</td>
</tr>
<tr>
<td>13</td>
<td>21.5</td>
<td>22.5</td>
<td>23.0</td>
<td>26.5</td>
</tr>
<tr>
<td>14</td>
<td>23.0</td>
<td>22.5</td>
<td>24.0</td>
<td>27.5</td>
</tr>
<tr>
<td>15</td>
<td>25.5</td>
<td>27.5</td>
<td>26.5</td>
<td>27.0</td>
</tr>
<tr>
<td>16</td>
<td>20.0</td>
<td>23.5</td>
<td>22.5</td>
<td>26.0</td>
</tr>
<tr>
<td>17</td>
<td>24.5</td>
<td>25.5</td>
<td>27.0</td>
<td>28.5</td>
</tr>
<tr>
<td>18</td>
<td>22.0</td>
<td>22.0</td>
<td>24.5</td>
<td>26.5</td>
</tr>
<tr>
<td>19</td>
<td>24.0</td>
<td>21.5</td>
<td>24.5</td>
<td>25.5</td>
</tr>
<tr>
<td>20</td>
<td>23.0</td>
<td>20.5</td>
<td>31.0</td>
<td>26.0</td>
</tr>
<tr>
<td>21</td>
<td>27.5</td>
<td>28.0</td>
<td>31.0</td>
<td>31.5</td>
</tr>
<tr>
<td>22</td>
<td>23.0</td>
<td>23.0</td>
<td>23.5</td>
<td>25.0</td>
</tr>
<tr>
<td>23</td>
<td>21.5</td>
<td>23.5</td>
<td>24.0</td>
<td>28.0</td>
</tr>
<tr>
<td>24</td>
<td>17.0</td>
<td>24.5</td>
<td>26.0</td>
<td>29.5</td>
</tr>
</tbody>
</table>

---

\(^1\)Potthoff and Roy (1964), *Biometrika*, pages 313-326.

\(^2\)Not sure what that is. Is it the same as the pterygomaxillary fissure? You can find pictures of that on the web.
Figure 1.1: Mouth size growth curves

Figure 1.1 contains a plot of each kid’s mouth size over time, where the girls are in red, and the boys are in blue. Generally, the sizes increase over time, and the boys tend to have larger mouths than the girls.

One approach to such growth curves is to assume that in the population, the growth is quadratic, where the coefficients are possibly different for girls and boys. That is, for $x = 8, 10, 12, 14$, the population average growth curves are, respectively,

$$f_{\text{Girls}}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 \quad \text{and} \quad f_{\text{Boys}}(x) = \beta_0 + \beta_1 x + \beta_2 x^2. \quad (1.1)$$
Now each individual’s values are going to deviate from the average for that one’s sex. The trick is to model these deviations. Generally, one would expect a person who is larger than average at age 8 will also be larger than average at age 10, etc., so one cannot assume all $27 \times 4$ deviations are independent. What kind of dependencies are there? Are the variances the same for the various ages? For the two sexes?

Some questions which can be addressed include

- Are cubic terms necessary, rather than just quadratic?
- Are the quadratic terms necessary (is $\alpha_2 = \beta_2 = 0$, so that the curves are straight lines)?
- Are the girls’ and boys’ curves the same (are $\alpha_j = \beta_j$ for $j = 0, 1, 2$)?
- Are the girls’ and boys’ curves parallel (are $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$, but maybe $\alpha_0 \neq \beta_0$)?
Part I

Distribution Theory
Chapter 2

Distributions and Densities

2.1 Probability models

Starting with the very general, suppose \( X \) is a random object. It could be a single variable, a vector, a matrix, or something more complicated, e.g., a function, infinite sequence, or image. The space of \( X \) is \( \mathcal{X} \), the set of possible values \( X \) can take on. Then a probability distribution on \( X \), or on \( \mathcal{X} \), is a function \( P \) that assigns a value in \([0, 1]\) to subsets of \( \mathcal{X} \). For “any” subset \( A \subset \mathcal{X} \), \( P[A] \) is the probability \( X \in A \). It can also be written \( P[X \in A] \). (The quotes on “any” are to point out that technically, only subsets in a “sigma field” of subsets of \( \mathcal{X} \) are allowed. We will gloss over that restriction, not because it is unimportant, but because for our purposes we do not get into too much trouble doing so.)

In order for \( P \) to be a probability distribution, it has to satisfy some axioms:

1. \( P[\mathcal{X}] = 1 \).

2. If \( A_1, A_2, \ldots \) are disjoint (\( A_j \cap A_j = \emptyset \) for \( i \neq j \)), then \( P[\bigcup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} P[A_i] \).

The second axiom means to refer to finite unions as well as infinite ones. Using these axioms, along with the restriction that \( 0 \leq P[A] \leq 1 \), all the usual properties of probabilities can be derived. Some such follow.

- **Complement.** The complement of a set \( A \) is \( A^c = \mathcal{X} - A \), that is, everything that is not in \( A \) (but in \( \mathcal{X} \)). Clearly, \( A \) and \( A^c \) are disjoint, and their union is everything:

  \[
  A \cap A^c = \emptyset, \quad A \cup A^c = \mathcal{X},
  \]

  so,

  \[
  1 = P[\mathcal{X}] = P[A \cup A^c] = P[A] + P[A^c],
  \]

  which means

  \[
  P[A^c] = 1 - P[A].
  \]

  That is, the probability the object does not landing \( A \) is 1 minus the probability that it does land in \( A \).

- \( P[\emptyset] = 0 \), because the empty set is the complement of \( \mathcal{X} \), which has probability 1.
• Union of two (nondisjoint) sets. If \( A \) and \( B \) are not disjoint, then it is not necessarily true that \( P[A \cup B] = P[A] + P[B] \). But \( A \cup B \) can be separated into two disjoint sets: The set \( A \) and the part of \( B \) not in \( A \), which is \( [B \cap A^c] \). Then

\[
P[A \cup B] = P[A] + P[B \cap A^c]. \tag{2.1}
\]

Now \( B = (B \cap A) \cup (B \cap A^c) \), and \( (B \cap A) \) and \( (B \cap A^c) \) are disjoint, so

\[
P[B] = P[B \cap A] + P[B \cap A^c] \Rightarrow P[B \cap A^c] = P[B] - P[A \cap B].
\]

Then stick that formula into (2.1), so that

\[
P[A \cup B] = P[A] + P[B] - P[A \cap B].
\]

The above definition doesn’t help much in specifying a probability distribution. In principle, one would have to give the probability of every possible subset, but luckily there are simplifications.

We will deal primarily with random variables and finite collections of random variables. A random variable has space \( \mathcal{X} \subset \mathbb{R} \), the real line. A collection of \( p \) random variables has space \( \mathcal{X} \subset \mathbb{R}^p \), the \( p \)-dimensional Euclidean space. The elements are usually arranged in some convenient way, such as in a vector (row or column), matrix, multidimensional array, or triangular array. Mostly, we will have them arranged as a column vector

\[
X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \tag{2.2}
\]

or a row vector.

Some common ways to specify the probabilities of a collection of \( p \) random variables include

1. Distribution functions
2. Densities
3. Moment generating functions (or characteristic functions)
4. Representations

Distribution functions and characteristic functions always exist, moment generating functions do not. The densities we will deal with are only those with respect to Lebesgue measure, or counting measure — which means for us, densities do not always exist. By “representation” we mean the random variables are expressed as a function of some other known random variables.

### 2.2 Distribution functions

The distribution function for \( X \) is the function \( F : \mathbb{R}^p \to [0,1] \) given by

\[
F(x) = F(x_1, \ldots, x_p) = P[X_1 \leq x_1, \ldots, X_p \leq x_p]. \tag{2.3}
\]

Note that \( F \) is defined on all of \( \mathbb{R}^p \), not just the space \( \mathcal{X} \). In principal, given \( F \), one can figure out the probability of all subsets \( A \subset \mathcal{X} \) (although no one would try), which
2.3. Densities

A density with respect to Lebesgue measure on \( \mathbb{R}^p \), which we simplify to “pdf” for “probability density function,” is a function \( f : \mathcal{X} \to [0, \infty) \) such that for any subset \( A \subset \mathcal{X} \),

\[
P[A] = \int \cdots \int_A f(x_1, x_2, \ldots, x_p) \, dx_1 \, dx_2 \cdots dx_p.
\]  

(2.5)
Chapter 2. Distributions and Densities

If $X$ has a pdf, then it is continuous. In fact, it is differentiable, $f$ being the derivative of $F$:

$$f(x_1, \ldots, x_p) = \frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_p} F(x_1, \ldots, x_p).$$

(2.6)

There are continuous distributions that do not have pdf’s, but they are weird. Any pdf has to satisfy the two properties,

1. $f(x_1, \ldots, x_p) \geq 0$ for all $(x_1, \ldots, x_p) \in \mathcal{X}$;

2. $\int \int \cdots \int_{\mathcal{X}} f(x_1, x_2, \ldots, x_p) dx_1 dx_2 \ldots dx_p = 1$.

It is also true that any function $f$ satisfying those two conditions is a pdf of a legitimate probability distribution.

Here is a list of some famous univariate (so that $p = 1$ and $\mathcal{X} \subseteq \mathbb{R}$) distributions with pdf’s:
2.3. Densities

<table>
<thead>
<tr>
<th>Name</th>
<th>Parameters</th>
<th>Space $\mathcal{X}$</th>
<th>pdf $f(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal: $N(\mu, \sigma^2)$</td>
<td>$\mu \in \mathbb{R}, \sigma^2 &gt; 0$</td>
<td>$\mathbb{R}$</td>
<td>$\frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$</td>
</tr>
<tr>
<td>Uniform: $U(a, b)$</td>
<td>$-\infty &lt; a &lt; b &lt; \infty$</td>
<td>$(a, b)$</td>
<td>$\frac{1}{b-a}$</td>
</tr>
<tr>
<td>Exponential: $\text{Exponential}(\lambda)$</td>
<td>$\lambda &gt; 0$</td>
<td>$(0, \infty)$</td>
<td>$\lambda \exp(-\lambda x)$</td>
</tr>
<tr>
<td>Gamma: $\text{Gamma}(\alpha, \lambda)$</td>
<td>$\alpha &gt; 0, \lambda &gt; 0$</td>
<td>$(0, \infty)$</td>
<td>$\frac{\lambda^\alpha}{\Gamma(\alpha)} \exp(-\lambda x) x^{\alpha-1}$</td>
</tr>
<tr>
<td>Beta: $\text{Beta}(\alpha, \beta)$</td>
<td>$\alpha &gt; 0, \beta &gt; 0$</td>
<td>$(0, 1)$</td>
<td>$\frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$</td>
</tr>
<tr>
<td>Cauchy</td>
<td></td>
<td>$\mathbb{R}$</td>
<td>$\frac{1}{\pi (1+x^2)}$</td>
</tr>
</tbody>
</table>

The $\Gamma(\alpha)$ is the Gamma function, presented later. There are many more, such as Student’s $t$, chi-squared, $F$, .... The most famous multivariate distribution is the multivariate normal. We will look at that one in Chapter 8.

**Example: A bivariate pdf**

Suppose $(X, Y)$ has space

$$\mathcal{W} = (0, 1) \times (0, 1) = \{(x, y) \mid 0 < x < 1, 0 < y < 1\}$$

and pdf

$$f(x, y) = c(x + y).$$

The constant $c$ is whatever it needs to be so that the pdf integrates to 1, i.e.,

$$1 = c \int_0^1 \int_0^1 (x + y)dydx = c \int_0^1 (x + 1/2)dx = c(1/2 + 1/2) = c.$$  \(2.9\)

So the pdf is simply $f(x, y) = x + y$. Some values of the distribution function are

$$F(0, 0) = 0$$

$$F\left(\frac{1}{2}, \frac{1}{4}\right) = \int_0^{1/2} \int_0^{1/4} (x + y)dydx = \int_0^{1/2} \left(\frac{x}{4} + \frac{1}{32}\right)dx = \frac{1}{32} + \frac{1}{32} = \frac{1}{16}$$

$$F(1/2, 2) = \int_0^{1/2} \int_0^1 (x + y)dydx = \int_0^{1/2} (x + 1/2)dx = \frac{1}{8} + \frac{1}{4} = \frac{3}{8}$$

$$F(2, 1) = 1.$$  \(2.10\)
Other probabilities:

\[
P[X + Y \leq \frac{1}{2}] = \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2} - x} (x + y) dy dx
\]
\[
= \int_0^{\frac{1}{2}} \left( x \left( \frac{1}{2} - x \right) + \frac{1}{2} \left( \frac{1}{2} - x \right)^2 \right) dx
\]
\[
= \int_0^{\frac{1}{2}} \left( \frac{1}{8} - \frac{x^2}{2} \right) dx
\]
\[
= \frac{1}{16} - \frac{1}{2} \left( \frac{1}{2} \right)^3
\]
\[
= \frac{1}{24}, \quad (2.11)
\]

and for \(0 < y < 1\),

\[
P[Y \leq y] = \int_0^1 \int_0^y (x + w) dw dx = \int_0^1 (xy + \frac{y^2}{2}) dx = \frac{y}{2} + \frac{y^2}{2} = \frac{1}{2} y(1 + y), \quad (2.12)
\]

which is the distribution function of \(Y\), at least for \(0 < y < 1\).

### 2.3.2 Probability mass functions

A **discrete** random variable is one for which \(\mathcal{X}\) is a countable (which includes finite) set. Its probability can be given by its **probability mass function**, which we will call “pmf,” \(f : \mathcal{X} \to [0, 1]\) given by

\[
P(\{x_1, \ldots, x_p\}) = f(x_1, \ldots, x_p). \quad (2.13)
\]

The pmf gives the probabilities of the individual points.\(^1\) The probability of any subset \(A\) is the sum of the probabilities of the individual points in \(A\).

Some univariate discrete distributions:

\(^1\)Measure-theoretically, the pmf is the density with respect to counting measure on \(\mathcal{X}\).
2.4. Distributions without pdf’s or pmf’s

Distributions need not be either discrete or having a density with respect to Lebesgue measure. We present here some simple examples.

The distribution of a discrete random variable is purely a jump function, that is, it is flat except for jumps of height $f(x)$ at $x$ for each $x \in \mathcal{X}$.

The most famous multivariate discrete distribution is the multinomial, which we look at in Example 3.4.3.
2.4.1 Example: Late start

Consider waiting for a train to leave. It will not leave early, but very well may leave late. There is a positive probability, say 10%, it will leave exactly on time. If it does not leave on time, there is a continuous distribution for how late it leaves. Thus it is not totally discrete, but not continuous, either, so it has neither a pdf nor pmf. It does have a distribution function, because everything does. A possible one is

\[
F(x) = \begin{cases} 
0 & \text{if } x < 0 \\
1/10 & \text{if } x = 0 \\
1 - (9/10) \exp(-x/100) & \text{if } x > 0,
\end{cases}
\]

(2.14)

where \( x \) is the number of minutes late. Here is a plot:

Is this a legitimate distribution function? It is easy to see it is nondecreasing, once one notes that \( 1 - (9/10) \exp(-x/100) > 1/10 \) if \( x > 0 \). The limits are ok as \( x \to \pm \infty \). It is also continuous from the right, where the only tricky spot is \( \lim_{x \downarrow 0} F(x) \), which goes to \( 1 - (9/10) \exp(0) = 1/10 = F(0) \), so it checks.

One can then find the probabilities of various late times, e.g., it has no chance of leaving early, 10% chance of leaving exactly on time, \( F(60) = 1 - (9/10) \exp(-60/100) \approx 0.506 \) chance of being at most one hour late, \( F(300) = 1 - (9/10) \exp(-300/100) \approx 0.955 \) chance of being at most five hours late, etc. (Sort of like Amtrak.)
2.4.2 Example: Spinner

Imagine a spinner whose pointer is one unit in length. It is spun so that it is equally likely to be pointing in any direction. The random quantity is the \((x, y)\) location of the end of the pointer, so that \(X = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}\), the circle with radius 1. The distribution of \((X, Y)\) is not discrete, because it can land anywhere on the circle, but it does not have a density with respect to Lebesgue measure on \(\mathbb{R}^2\) because the integral over the circle is the volume above the circle under the pdf, that volume being 0.

But there is a distribution function. The \(F(x, y)\) is the arc length of the part of the circle that has \(x\)-coordinate less than or equal to \(x\) and \(y\)-coordinate less than or equal to \(y\), divided by total arc length (which is \(2\pi\)):

\[
F(x, y) = \frac{\text{arc length}(\{(u, v) \mid u^2 + v^2 = 1, u \leq x, v \leq y\})}{2\pi}.
\] (2.15)

Here is a sketch of \(F\):

But there is an easier way to describe the distribution. For any point \((x, y)\) on the circle, one can find the angle with the \(x\)-axis of the line connecting \((0, 0)\) and \((x, y)\), \(\theta = \text{Angle}(x, y)\), so that \(x = \cos(\theta)\) and \(y = \sin(\theta)\). For uniqueness’ sake, take
\( \theta \in [0, 2\pi) \). Then \((x, y)\) being uniform on the circle implies that \(\theta\) is uniform from 0 to \(2\pi\). Then the distribution of \((X, Y)\) can be described via

\[
(X, Y) = (\cos(\theta), \sin(\theta)), \text{ where } \theta \sim \text{Uniform}[0, 2\pi].
\] (2.16)

Such a description is called a representation, in that we are representing one set of random variables as a function of another set (which in this case is just the one \(\theta\)).

### 2.4.3 Example: Random coin

Imagine now a two-stage process, where one first chooses a coin out of an infinite collection of coins, then flips the coin \(n = 10\) times independently. The coins have different probabilities of heads \(x\), so that over the population of coins, \(X \sim \text{Uniform}(0, 1)\). Let \(Y\) be the number of heads among the \(n\) flips. Then what is random is \((X, Y)\). This vector is neither discrete nor continuous: \(X\) is continuous and \(Y\) is discrete. The space is a union of 11 (= \(n + 1\)) line segments,

\[
\mathcal{W} = \{(x, 0) \mid 0 < x < 1\} \cup \{(x, 1) \mid 0 < x < 1\} \cup \cdots \cup \{(x, 10) \mid 0 < x < 1\}. \quad (2.17)
\]

Here is the graph of \(\mathcal{W}\):
Chapter 3

Expected Values

3.1 Definition

The distribution function $F$ contains all there is to know about the distribution of a random vector, but it is often difficult to take in all at once. Quantities that summarize aspects of the distribution are often helpful, many of which are expected values of functions of $X$. We define expected value in the pdf and pmf cases. There are many $X$’s that have neither a pmf nor pdf, but even in those cases we can often find the expected value.

**Definition 1. Expected Value.** Suppose $X$ has pdf $f$, and $g : \mathcal{X} \to \mathbb{R}$. If

$$
\int \cdots \int_{\mathcal{X}} |g(x_1, \ldots, x_p)| f(x_1, x_2, \ldots, x_p) \, dx_1 \cdots dx_p < \infty,
$$

(3.1)

then the expected value of $g(X)$, $E[g(X)]$, exists and

$$
E[g(X)] = \int \cdots \int_{\mathcal{X}} g(x_1, \ldots, x_p) f(x_1, x_2, \ldots, x_p) \, dx_1 \cdots dx_p.
$$

(3.2)

If $X$ has pmf $f$, and

$$
\sum \cdots \sum_{(x_1, \ldots, x_p) \in \mathcal{X}} |g(x_1, \ldots, x_p)| f(x_1, x_2, \ldots, x_p) < \infty,
$$

(3.3)

then the expected value of $g(X)$, $E[g(X)]$, exists and

$$
E[g(X)] = \sum \cdots \sum_{(x_1, \ldots, x_p) \in \mathcal{X}} g(x_1, \ldots, x_p) f(x_1, x_2, \ldots, x_p).
$$

(3.4)

The requirement (3.1) that the absolute value of the function must have a finite integral is there to eliminate ambiguous situations. For example, consider the Cauchy distribution with pdf $f(x) = (1/\pi)(1/(1 + x^2))$ and space $\mathbb{R}$, and take $g(x) = x$, so we wish to find $E[X]$. Consider

$$
\int_{-\infty}^{\infty} |x| f(x) \, dx = \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{|x|}{1 + x^2} \, dx = 2 \int_{0}^{\infty} \frac{1}{\pi} \frac{x}{1 + x^2} \, dx.
$$

(3.5)
For large $|x|$, the integrand is on the order of $1/|x|$, which does not have a finite integral. More precisely, it is not hard to show that

$$\frac{x}{1 + x^2} > \frac{1}{2x} \text{ for } x > 1.$$  

Thus

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \frac{|x|}{1 + x^2} dx > 2 \int_1^\infty \frac{1}{\pi} \frac{1}{x} dx = \left(\frac{2}{\pi}\log(x)\right|_1^\infty = \left(\frac{2}{\pi}\log(\infty) = \infty. \right.$$  

Thus, according to the definition, we say that “the expected value of the Cauchy does not exist.” By the symmetry of the density, it would be natural to expect the expected value to be 0. But, what we have is

$$E[X] = \int_{-\infty}^{0} xf(x) dx + \int_{0}^{\infty} xf(x) dx = -\infty + \infty = \text{Undefined.}$$  

That is, we cannot do the integral, so the expected value is not defined.

One could allow $+\infty$ and $-\infty$ to be legitimate values of the expected value, e.g., say that $E[X^2] = +\infty$ for the Cauchy, as long as the value is unambiguous. We are not allowing that possibility formally, but informally will on occasion act as though we do.

Expected values cohere in the proper way, that is, if $Y$ is a random vector that is a function of $X$, say $Y = h(X)$, then for a function $g$ of $Y$,

$$E[g(Y)] = E[g(h(X))],$$  

if the latter exists. This property helps in finding the expected values when representations are used. For example, in the spinner case (2.16),

$$E[X] = E[\cos(\theta)] = \frac{1}{2\pi} \int_0^{2\pi} \cos(\theta) d\theta = 0,$$  

where the first expected value has $X$ as the random variable, for which we do not have a pdf, and the second expected value has $\theta$ as the random variable, for which we do have a pdf (the Uniform$[0, 2\pi]$).

An important feature of expected values is their linearity, which follows by the linearity of integrals and sums:

**Lemma 1.** For any random variables $X, Y$, and constant $c$,

$$E[cX] = cE[X] \text{ and } E[X + Y] = E[X] + E[Y],$$  

if the expected values exist.

The lemma can be used to show more involved linearities, e.g.,


(since $E[d] = d$ for a constant $d$), and

$$E[g(X) + h(X)] = E[g(X)] + E[h(X)].$$  

3.2 Means, Variances, and Covariances

Warning. Be aware that for non-linear functions, the expected value of a function is NOT the function of the expected value, i.e.,

$$E[g(X)] \neq g(E[X])$$

(3.14)

unless $g(x)$ is linear, or you are lucky. For example,

$$E[X^2] \neq (E[X])^2,$$

(3.15)

unless $X$ is a constant. (Which is fortunate, because otherwise all variances would be 0. See equation 3.20 below.)

3.1.1 Indicator functions

An indicator function is one which takes on only the values 0 and 1. It is usually given as $I_A$ for a subset $A \subset \mathcal{X}$, where $A$ contains the values for which the function is 1:

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} .$$

(3.16)

These functions give an alternative expression for probabilities in terms of expected values:

$$E[I_A(X)] = 1 \times P[X \in A] + 0 \times P[X \notin A] = P[A].$$

(3.17)

3.2 Means, Variances, and Covariances

Means, variances, and covariances are particular expected values. For a random variable, the mean is just its expected value:

The mean of $X = E[X]$ (often denoted $\mu$). (3.18)

(From now on, we will usually suppress the phrase “if it exists” when writing expected values, but think of it to yourself when reading “$E$.”) The variance is the expected value of the deviation from the mean, squared:

The variance of $X = \text{Var}[X] = E[(X - E[X])^2]$ (often denoted $\sigma^2$). (3.19)

The standard deviation is the square root of the variance. It is often a nicer quantity because it is in the same units as $X$, and measures the “typical” size of the deviation of $X$ from its mean.

A very useful formula for finding variances is

$$\text{Var}[X] = E[X^2] - (E[X])^2,$$

(3.20)

which can be seen, letting $\mu = E[X]$, as follows:

$$E[(X - \mu)^2] = E[X^2 - 2X\mu + \mu^2] = E[X^2] - 2E[X]\mu + \mu^2 = E[X^2] - \mu^2.$$  

(3.21)

With two random variables, $(X, Y)$, say, there is in addition the covariance:

The covariance of $X$ and $Y = \text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])].$  

(3.22)
The covariance measures a type of relationship between \( X \) and \( Y \). Notice that the expectand is positive when \( X \) and \( Y \) are both greater than or both less than their respective means, and negative is one is greater and one less. Thus if \( X \) and \( Y \) tend to go up or down together, the covariance will be positive, while if when one goes up the other goes down, the covariance will be negative. Note also that it is symmetric, \( \text{Cov}[X, Y] = \text{Cov}[Y, X] \), and \( \text{Cov}[X, X] = \text{Var}[X] \).

As for the variance in (3.20), we have the formula

\[
\]

The correlation coefficient is a normalization of the covariance, which is generally easier to interpret:

The correlation coefficient of \( X \) and \( Y \) is

\[
\text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}.
\]

This is a unitless quantity that measures the linear relationship of \( X \) and \( Y \). It is bounded by \(-1\) and \(+1\). To see this, we first need the following.

Lemma 2. Cauchy-Schwarz. For random variables \((U, V)\),

\[
E[UV]^2 \leq E[U^2]E[V^2],
\]

with equality if and only if

\[
U = 0 \text{ or } V = \beta U \text{ with probability } 1,
\]

for \( \beta = E[UV]/E[U^2] \).

Here, the phrase “with probability 1” means \( P[U = 0] = 1 \) or \( P[V = aU] = 1 \). It is a technicality that you can usually ignore. It is there because for continuous variables, for example, you can always change a few points (which have 0 probability) without affecting the distribution.

Proof. The lemma is easy to see if \( U \) is always 0, because then \( E[UV] = E[U^2] = 0 \). Suppose it is not, so that \( E[U^2] > 0 \). Consider

\[
\]

Because the expectand on the left is nonnegative, so is the expected value for any \( b \). In particular, it is nonnegative for the \( b \) that minimizes the expected value, which is easy to find:

\[
\frac{\partial}{\partial b} E[(V - bU)^2] = -2E[UV] + 2bE[U^2],
\]

and setting that to 0 yields \( b = \beta \) where

\[
\beta = \frac{E[UV]}{E[U^2]}.
\]

Then

\[
E[V^2] - 2\beta E[UV] + \beta^2 E[U^2] = E[V^2] - 2\frac{E[UV]}{E[U^2]} E[UV] + \left(\frac{E[UV]}{E[U^2]}\right)^2 E[U^2]
\]

\[
= E[V^2] - \frac{E[UV]^2}{E[U^2]}
\]

\[
\geq 0,
\]
3.2. Means, Variances, and Covariances

from which (3.25) follows.

There is equality in (3.25) if and only if there is equality in (3.30), which means that

\[ E[(V - \beta U)^2] = 0. \] (3.31)

Because the expectand is nonnegative, its expected value can be 0 if and only if it is 0, i.e.,

\[ (V - \beta U)^2 = 0 \text{ with probability 1.} \] (3.32)

But that equation implies the second part of (3.26), proving the lemma.

For variables \((X, Y)\), apply the lemma with \(U = X - E[X]\) and \(V = Y - E[Y]\):

\[
E[(X - E[X])(Y - E[Y])^2] \leq E[(X - E[X])^2]E[(Y - E[Y])^2]
\]

\[ \iff \quad \text{Cov}[X, Y]^2 \leq \text{Var}[X]\text{Var}[Y]. \] (3.33)

Thus from (3.24),

\[ -1 \leq \text{Corr}[X, Y] \leq 1. \] (3.34)

Furthermore, if there is an equality in (3.33), then either \(X\) is a constant, or

\[ Y - E[Y] = b(X - E[X]) \iff Y = \alpha + \beta X, \] (3.35)

where

\[ \beta = \frac{\text{Cov}[X, Y]}{\text{Var}[X]} \text{ and } \alpha = E[Y] - \beta E[X]. \] (3.36)

In this case,

\[ \text{Corr}[X, Y] = \begin{cases} 
1 & \text{if } \beta > 0 \\
-1 & \text{if } \beta < 0
\end{cases}. \] (3.37)

Thus the correlation coefficient measures the linearity of the relationship between \(X\) and \(Y\), +1 meaning perfectly positively linearly related, −1 meaning perfectly negatively linearly related.

3.2.1 Example

Suppose \((X, Y)\) has pdf \(f(x, y) = 2\) for \((x, y) \in \mathcal{W} = \{(x, y) \mid 0 < x < y < 1\}\), which is the upper-left triangle of the unit square:
One would expect the correlation to be positive, since the large y’s tend to go with larger x’s, but the correlation would not be +1, because the space is not simply a line segment. To find the correlation, we need to perform some integrals:

\[
E[X] = \int_0^1 \int_0^y x \, dx \, dy = \int_0^1 y^2 \, dy = \frac{1}{3}, \quad E[Y] = \int_0^1 \int_0^y y \, dx \, dy = 2 \int_0^1 y^2 \, dy = \frac{2}{3},
\]

(3.38)

\[
E[X^2] = \int_0^1 \int_0^y x^2 \, dx \, dy = \int_0^1 \frac{2y^3}{3} \, dy = \frac{1}{6}, \quad E[Y^2] = \int_0^1 \int_0^y y^2 \, dx \, dy = 2 \int_0^1 y^3 \, dy = \frac{1}{3},
\]

(3.39)

and

\[
E[XY] = \int_0^1 \int_0^y xy \, dx \, dy = \int_0^1 y^3 \, dy = \frac{1}{4}.
\]

(3.40)

Then

\[
Var[X] = \frac{1}{6} - \frac{1}{3^2} = \frac{1}{18}, \quad Var[Y] = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}, \quad Cov[X, Y] = \frac{1}{4} - \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{36},
\]

(3.41)

and, finally,

\[
Corr[X, Y] = \frac{\frac{1}{36}}{\sqrt{\frac{1}{18} \cdot \frac{1}{18}}} = \frac{1}{2}.
\]

(3.42)

This value does seem plausible: positive but not too close to 1.

### 3.2.2 The variance of linear combinations & affine transformations

A linear combination of the variables, \(X_1, \ldots, X_p\), is a function of the form

\[
b_1 X_1 + \cdots + b_p X_p,
\]

(3.43)
An affine transformation just adds a constant:
\[ a + b_1 X_1 + \cdots + b_p X_p. \] (3.44)

Thus they are almost the same, and if you want to add the (constant) variable \( X_0 \equiv 1 \), you can think of an affine transformation as a linear combination, as one does when setting up a linear regression model with intercept. Here we find formulas for the variance of an affine transformation.

Start with \( a + bX \):
\[
\begin{align*}
\text{Var}[a + bX] &= E[(a + bX - E[a + bX])^2] \\
&= E[(a + bX - a - bE[X])^2] \\
&= E[b^2(X - E[X])^2] \\
&= b^2E[(X - E[X])^2] \\
&= b^2\text{Var}[X].
\end{align*}
\] (3.45)

The constant \( a \) goes away (it does not contribute to the variability), and the constant \( b \) is squared. For a linear combination of two variables, the variance involves the two variances, as well as the covariance:
\[
\begin{align*}
\text{Var}[a + bX + cY] &= E[(a + bX + cY - E[a + bX + cY])^2] \\
&= E[(b(X - E[X]) + c(Y - E[Y]))^2] \\
&= b^2E[(X - E[X])^2] + 2bc E[(X - E[X])(Y - E[Y])] + c^2E[(Y - E[Y])^2] \\
&= b^2\text{Var}[X] + c^2\text{Var}[Y] + 2bc \text{Cov}[X,Y].
\end{align*}
\] (3.46)

With \( p \) variables, we have
\[
\text{Var}[a + \sum_{i=1}^{p} b_i X_i] = \sum_{i=1}^{p} b_i^2 \text{Var}[X_i] + 2 \sum_{1 \leq i < j \leq p} b_i b_j \text{Cov}[X_i, X_j].
\] (3.47)

This formula can be made simpler using matrix and vector notation, which we do in the next section.

### 3.3 Other summaries

Means, variances and covariances are often good summaries of distributions; others are mentioned below. These summaries typically do not characterize the distributions. For example, two random variables could have the same means and same variances, but still be quite different: a Binomial(16, 1/2) has mean 8 and variance 4, as does a Normal(8, 4). The distribution function contains all the information about a random variable, or a random sample, as does the probability mass function or probability density function, if such exists, as well as the moment generating function.

Some types of summaries:

- **Centers.** A measure of center is a number that is “typical” of the entire distribution. The mean and the median are the two most popular, but there are also trimmed means, the tri-mean, the mode, and others.
• Position – Quantiles. A positional measure is one that gives the value that is in a certain relation to the rest of the values. For example, the .25 quantile is the value such that the random variable is (approximately) below the value 25% of the time, and above it 75% of the time. The median is the 1/2 quantile. More precisely, for \( q \in [0, 1] \), a \( q^{th} \) quantile is any value \( \eta_q \) such that

\[
P[X \leq \eta_q] \geq q \quad \text{and} \quad P[X \geq \eta_q] \geq 1 - q.
\]

A particular quantile may not be unique. It is if the distribution function is strictly increasing at the value.

• Spreads. A measure of spread indicates how spread out the distribution is, that is, how much do the values vary from the center. Famous ones are the standard deviation (and variance); interquartile range, which is the difference between the \( 3/4^{th} \) quantile and \( (1/4)^{th} \) quantile; the range, which is the maximum minus the minimum; and the MAD, median absolute deviation, \( E[|X - \eta|] \), where \( \eta \) is the median of \( X \).

• Moments. The \( k^{th} \) raw moment of a random variable is the expected value of its \( k^{th} \) power, where \( k = 1, 2, \ldots \). The \( k^{th} \) central moment is the expected value of the \( k^{th} \) power of its deviation from the mean \( \mu \), at least for \( k > 1 \):

\[
\begin{align*}
  k^{th} \text{ raw moment} & = \mu'_k = E[X^k], \quad k = 1, 2, \ldots \\
  k^{th} \text{ central moment} & = \mu_k = E[(X - \mu)^k], \quad k = 2, 3, \ldots .
\end{align*}
\]

Thus \( \mu'_1 = \mu = E[X] \), \( \mu'_2 = E[X^2] \), and \( \mu_2 = \sigma^2 = \text{Var}[X] = \mu'_2 - \mu^2_1 \). The first central moment \( \mu_1 \) is \( \mu'_1 = \mu \), the mean. It is not hard, but a bit tedious, to figure out the \( k^{th} \) central moment from the first \( k \) raw moments, and \textit{vice versa}. It is not uncommon for given moments not to exist. In particular, if the \( k^{th} \) moment does not exist, then neither does any higher moment.

The third central moment is generally measure of \textit{skewness}, where symmetric distributions have 0 skewness, and a tail heavier to the right than to the left would have a positive skewness. Usually it is normalized so that it is not dependent on the variance:

\[
\text{Skewness} = \kappa_3 = \frac{\mu'_3}{\sigma^3}.
\]

The fourth central moment is a measure of \textit{kurtosis}. It, too, is normalized:

\[
\text{Kurtosis} = \kappa_4 = \frac{\mu'_4}{\sigma^4} - 3 .
\]

The \textit{Normal}(0, 1) has \( \mu_4 = 3 \), so that subtracting 3 in (3.51) means the kurtosis of a normal is 0. It is not particularly easy to figure out what kurtosis means in general, but for nice unimodal densities, it measures “boxiness.” A negative kurtosis indicates a density more boxy than the Normal. The Uniform is boxy. A positive kurtosis indicates a pointy middle and heavy tails, such as the double exponential.

• Generating Functions. A generating function is a meta-summary. Typically, derivatives of the generating function yield certain quantities, e.g., the moment generating function generates moments, the cumulant generating function generates cumulants, etc.
3.4 The moment and cumulant generating functions

The moment generating function (mgf for short) of $X$ is a function from $\mathbb{R}^p \to [0, \infty]$ given by

$$M_X(t_1, \ldots, t_p) = E \left[ e^{t_1 X_1 + \cdots + t_p X_p} \right].$$ (3.52)

The mgf does not always exist, that is, often the integral or sum defining the expected value diverges. That is ok, as long as it is finite for $t = (t_1, \ldots, t_p)$ in a neighborhood of $\mathbb{Q}_p$. If that property holds, then the mgf uniquely determines the distribution of $X$.

**Theorem 1. Uniqueness of MGF** If for some $\epsilon > 0$,

$$M_X(t) < \infty \text{ and } M_X(t) = M_Y(t) \text{ for all } t \text{ such that } \|t\| \leq \epsilon,$$ (3.53)

then $X$ and $Y$ have the same distribution.

If one knows complex variables, the characteristic function is superior because it always exists. It is defined as $\phi_X(t) = E[\exp(i(t_1 X_1 + \cdots + t_N X_N))]$.

The uniqueness in Theorem 1 is the most useful property of mgf’s, but they can also be useful in generating moments, as in (3.57). We next extend the result to vectors.

**Lemma 3.** Suppose $X$ has mgf such that for some $\epsilon > 0$,

$$M_X(t) < \infty \text{ for all } t \text{ such that } \|t\| \leq \epsilon.$$ (3.54)

Then for any nonnegative integers $k_1, \ldots, k_p$,

$$E[X_1^{k_1} X_2^{k_2} \cdots X_p^{k_p}] = \frac{\partial^{k_1+\cdots+k_p}}{\partial t_1^{k_1} \cdots \partial t_p^{k_p}} M_X(t) \bigg|_{t=0},$$ (3.55)

which is finite.

The quantities $E[X_1^{k_1} X_2^{k_2} \cdots X_p^{k_p}]$ are called mixed moments. The most common one is just $E[X_1 X_2]$, used in finding the covariance. Notice that this lemma implies that all mixed moments are finite under the condition (3.54). This condition is actually quite strong.

Specializing to a random variable $X$, the mgf is

$$M_X(t) = E[e^{tX}].$$ (3.56)

If it exists for $t$ in a neighborhood of 0, then all moments of $X$ exist, and

$$\frac{\partial^k}{\partial t^k} M_X(t) \bigg|_{t=0} = E[X^k].$$ (3.57)

The cumulant generating function is the log of the moment generating function,

$$c_X(t) = \log(M_X(t)).$$ (3.58)

It generates the cumulants, which are defined by what the cumulant generating function generates, i.e., for a random variable, the $k^{th}$ cumulant is

$$\gamma_k = \frac{\partial^k}{\partial t^k} c_X(t) \bigg|_{t=0}.$$ (3.59)
cumulants are slightly easier to work with than moments. The first four are

\[ \gamma_1 = E[X] = \mu_1 = \mu \]
\[ \gamma_2 = Var[X] = \mu_2 = \sigma^2 \]
\[ \gamma_3 = E[(X - E[X])^3] = \mu_3 \]
\[ \gamma_4 = E[(X - E[X])^4] - 3 \text{Var}[X]^2 = \mu_4 - 3\mu_2^2 = \mu_4 - 3\sigma^4. \quad (3.60) \]

The skewness (3.50) and kurtosis (3.51) are then simple functions of the cumulants:

\[ \text{Skewness}[X] = \frac{\gamma_3}{\sigma^3} \quad \text{and} \quad \text{Kurtosis}[X] = \frac{\gamma_4}{\sigma^4}. \quad (3.61) \]

### 3.4.1 Example: Normal

Suppose \( X \sim N(\mu, \sigma^2) \) (with \( \sigma^2 > 0 \)). Then its mgf is

\[
M_X(t) = E[e^{tX}] = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{1}{2\sigma^2} (x-\mu)^2} dx
\]

\[
= \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} e^{tx-\frac{1}{2} \frac{1}{\sigma^2} (x^2-2\mu x+\mu^2)}
\]

\[
= \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} e^{-\frac{1}{2} \frac{1}{\sigma^2} (x^2-2(\mu+tv^2)x+\mu^2)}. \quad (3.62)
\]

In the exponent, complete the square with respect to the \( x \):

\[
x^2 - 2(\mu + tv^2)x = (x - (\mu + tv^2))^2 - (\mu + tv^2)^2. \quad (3.63)
\]

Then

\[
M_X(t) = e^{\frac{1}{2} \frac{1}{\sigma^2} ((\mu+tv^2)^2-\mu^2)} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2} \frac{1}{\sigma^2} (x-(\mu+tv^2))^2} dx. \quad (3.64)
\]

Notice that the integrand is the pdf of a \( N(\mu + tv^2, \sigma^2) \), which means the integral is 1. Expanding the \( (\mu + tv^2)^2 \) and simplifying a bit yields

\[
M_X(t) = e^{t\mu + \frac{1}{2} tv^2}, \quad t \in \mathbb{R}. \quad (3.65)
\]

The cumulant generating function is then a simple quadratic:

\[
c_X(t) = t\mu + \frac{1}{2} \sigma^2 t^2, \quad (3.66)
\]

and it is easy to see that

\[
c_X'(0) = \mu, \quad c_X''(0) = \sigma^2, \quad c_X'''(t) = 0. \quad (3.67)
\]

Thus the mean is \( \mu \) and variance is \( \sigma^2 \) (not surprisingly), and all other cumulants are 0. In particular, the skewness and kurtosis are both 0.
3.4.2 Example: Gamma

The Gamma distribution has two parameters: \( \alpha > 0 \) is the shape parameter, and \( \lambda > 0 \) is the scale parameter. Its space is \( \mathcal{X} = (0, \infty) \), and its pdf is \( c \, x^{\alpha-1} \exp(-\lambda x) \), where

\[
\frac{1}{c} = \int_0^\infty x^{\alpha-1} \exp(-\lambda x) \, dx \\
= \int_0^\infty (u/\lambda)^{\alpha-1} \exp(-u) \, du / \lambda \quad (u = \lambda x) \\
= \frac{1}{\lambda^\alpha} \int_0^\infty u^{\alpha-1} \exp(-u) \, du \\
= \frac{1}{\lambda^\alpha} \Gamma(\alpha). \tag{3.68}
\]

The gamma function is defined to be that very integral, i.e., for \( \alpha > 0 \),

\[
\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} \exp(-u) \, du. \tag{3.69}
\]

It generally has no closed-form value, but there are some convenient facts:

\[
\Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \quad \text{for} \quad \alpha > 0; \\
\Gamma(n) = (n-1)! \quad \text{for} \quad n = 1, 2, \ldots; \\
\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \tag{3.70}
\]

The pdf for a Gamma distribution is then

\[
f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \, x^{\alpha-1} \exp(-\lambda x), \quad x \in (0, \infty). \tag{3.71}
\]

If \( \alpha = 1 \), then this distribution is called the Exponential (\( \lambda \)), with pdf

\[
g(x \mid \lambda) = \lambda \, e^{-\lambda x}, \quad x \in \mathcal{X} = (0, \infty). \tag{3.72}
\]

The mgf is

\[
M_X(t) = E[e^{tX}] = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{tx} x^{\alpha-1} \exp(-\lambda x) \\
= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} \exp(-(\lambda-t)x). \tag{3.73}
\]

That integral needs \( (\lambda - t) > 0 \) to be finite, so we need \( t < \lambda \), which means the mgf is finite for a neighborhood of zero, since \( \lambda > 0 \). Now the integral at the end of (3.73) is the same as that in (3.68), except that \( \lambda - t \) is in place of \( \lambda \). Thus we have

\[
E[e^{tX}] = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\lambda-t)^\alpha} \\
= \left( \frac{\lambda}{\lambda - t} \right)^\alpha, \quad t < \lambda. \tag{3.74}
\]
We will use the cumulant generating function \( c_X(t) = \log(M_X(t)) \) to obtain the mean and variance, because it is slightly easier. Thus

\[
c'_X(t) = \frac{\partial}{\partial t} \alpha (\log(\lambda) - \log(\lambda - t)) = \frac{\alpha}{\lambda - t} \implies E[X] = c'_X(0) = \frac{\alpha}{\lambda}, \tag{3.75}
\]

and

\[
c''_X(t) = \frac{\partial^2}{\partial t^2} \alpha (\log(\lambda) - \log(\lambda - t)) = \frac{\alpha}{(\lambda - t)^2} \implies Var[X] = c''_X(0) = \frac{\alpha}{\lambda^2}. \tag{3.76}
\]

In general, the \( k^{th} \) cumulant (3.59) is

\[
\gamma_k = (k - 1)! \frac{\alpha}{\lambda^k}, \tag{3.77}
\]

and in particular

\[
Skewness[X] = \frac{2\alpha/\lambda^3}{\alpha^3/\lambda^3} = \frac{2}{\sqrt{\alpha}} \quad \text{and} \quad \frac{\alpha}{\lambda^4} \tag{3.78}
\]

Kurtosis[X] = \frac{6\alpha/\lambda^4}{\alpha^2/\lambda^4} = \frac{6}{\alpha}.

Thus the skewness and kurtosis depends on just the shape parameter \( \alpha \). Also, they are positive, but tend to 0 as \( \alpha \) increases.

### 3.4.3 Example: Binomial and Multinomial

The Binomial is a model for counting the number of successes in \( n \) trials, e.g., the number of heads in ten flips of a coin, where the trials are independent (formally defined in Section 4.3) and have the same probability \( p \) of success. As in Section 2.3.2,

\[
X \sim Binomial(n, p) \implies f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \mathcal{X} = \{0, 1, \ldots, n\}. \tag{3.79}
\]

The fact that the pdf sums to 1 relies on the binomial theorem:

\[
(a + b)^n = \sum_{x=0}^{n} a^x b^{n-x}, \tag{3.80}
\]

with \( a = p \) and \( b = 1 - p \). This theorem also helps in finding the mgf:

\[
M_X(t) = E[e^{tX}] = \sum_{x=0}^{n} e^{tx} f_X(x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = (pe^t + 1 - p)^n. \tag{3.81}
\]
3.4. The moment and cumulant generating functions

It is finite for all $t \in \mathbb{R}$, as is the case for any bounded random variable. The first two moments are

$$E[X] = M'_X(0) = n(pe^t + 1 - p)^{n-1}pe^t \bigg|_{t=0} = np,$$  

(3.82)

and

$$E[X^2] = M''_X(0) = [n(n-1)(pe^t + 1 - p)^{n-2}(pe^t)^2 + n(pe^t + 1 - p)^{n-1}pe^t] \bigg|_{t=0} = n(n-1)p^2 + np.$$  

(3.83)

Thus

$$\text{Var}[X] = E[X^2] - E[X]^2 = n(n-1)p^2 + np - (np)^2 = np - np^2 = np(1 - p).$$  

(3.84)

(In Section 5.3.4 we will exhibit an easier approach.)

The Multinomial distribution also models the results of $n$ trials, but here there are $K$ possible categories for each trial. E.g., one may roll a die $n$ times, and see whether it is a one, two, . . . , or six (so $K = 6$); or one may randomly choose $n$ people, each of whom is then classified as short, medium, or tall (so $K = 3$). As for the Binomial, the trials are assumed independent, and the probability of an individual trial coming up in category $k$ is $p_k$, so that $p_1 + \cdots + p_K = 1$. The random vector is $\mathbf{X} = (X_1, \ldots, X_K)'$, where $X_k$ is the number of observations from category $k$. Letting $\underline{p} = (p_1, \ldots, p_K)'$, we have

$$\mathbf{X} \sim \text{Multinomial}(n, \underline{p}) \implies f_{\mathbf{X}}(\mathbf{x}) = \binom{n}{\mathbf{x}} p_1^{x_1} \cdots p_K^{x_K}, \quad \mathbf{x} \in \mathcal{X},$$  

(3.85)

where the space consists of all possible ways $K$ nonnegative integers can sum to $n$:

$$\mathcal{X} = \{ \mathbf{x} \in \mathbb{R}^K \mid x_k \in \{0, \ldots, n\} \text{ for each } k, \text{ and } x_1 + \cdots + x_K = n \},$$  

(3.86)

and for $\mathbf{x} \in \mathcal{X}$,

$$\binom{n}{\mathbf{x}} = \frac{n!}{x_1! \cdots x_K!}.$$  

(3.87)

This pmf is related to the multinomial theorem:

$$(a_1 + \cdots + a_K)^n = \sum_{\mathbf{x} \in \mathcal{X}} \binom{n}{\mathbf{x}} a_1^{x_1} \cdots a_K^{x_K}.$$  

(3.88)

Note that the Binomial is basically a special case of the Multinomial with $K = 2$:

$$X \sim \text{Binomial}(n, p) \implies (X, n - X)' \sim \text{Multinomial}(n, (p, 1 - p)').$$  

(3.89)
Chapter 3. Expected Values

Now for the mgf. It is a function of \( t = (t_1, \ldots, t_K) \):
\[
M_X(t) = E[e^{t'X}] = \sum_{x \in X} e^{t'x} f_X(x) \\
= \sum_{x \in X} \left( \begin{array}{c} n \\ x \end{array} \right) (p_1 e^{t_1})^{x_1} \cdots (p_K e^{t_K})^{x_K} \\
= (p_1 e^{t_1} + \cdots + p_K e^{t_K})^n < \infty \text{ for all } t \in \mathbb{R}^K.
\] (3.90)

The mean and variance of each \( X_k \) can be found much as for the Binomial. We find that
\[
E[X_k] = np_k \text{ and } \text{Var}[X_k] = np_k(1-p_k). \tag{3.91}
\]

For the covariance between \( X_1 \) and \( X_2 \), we first find
\[
E[X_1 X_2] = \frac{\partial^2}{\partial t_1 \partial t_2} M_X(t) \bigg|_{t=\mathbf{0}_k} \\
= n(n-1)(p_1 e^{t_1} + \cdots + p_K e^{t_K})^{n-2} p_1 e^{t_1} p_2 e^{t_2} \bigg|_{t=\mathbf{0}_k} \\
= n(n-1)p_1 p_2. \tag{3.92}
\]

Thus
\[
\text{Cov}[X_1, X_2] = n(n-1)p_1 p_2 - (np_1)(np_2) = -n p_1 p_2. \tag{3.93}
\]

Similarly, \( \text{Cov}[X_k, X_l] = -n p_k p_l \) is \( k \neq l \). It does make sense for the covariance to be negative, since the more there are in category 1, the fewer are available for category 2.

3.5 Vectors and matrices

The mean of a vector or matrix of random variables is the corresponding vector or matrix of means. That is, if \( X \) is a \( p \times 1 \) vector, \( X = (X_1, \ldots, X_p)' \), then
\[
E[X] = \left( \begin{array}{c} E[X_1] \\ \vdots \\ E[X_p] \end{array} \right). \tag{3.94}
\]

Similarly, if \( X \) is an \( n \times p \) matrix, then so is its mean:
\[
E[X] = E \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} E(X_{11}) & E(X_{12}) & \cdots & E(X_{1p}) \\ E(X_{21}) & E(X_{22}) & \cdots & E(X_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{n1}) & E(X_{n2}) & \cdots & E(X_{np}) \end{bmatrix}. \tag{3.95}
\]

The linearity in Lemma 1 holds for vectors and matrices as well. That is, for fixed \( q \times p \) matrix \( B \) and \( q \times 1 \) vector \( \mathbf{a} \),
\[
E[\mathbf{a} + B X] = \mathbf{a} + BE[X], \tag{3.96}
\]
and for matrices $A (m \times q)$, $B (m \times n)$ and $C (p \times q)$,

$$A + E[BXC] = A + BE[X]C.$$  \hfill (3.97)

These formulas can be proved by writing out the individual elements, and noting that each is a linear combination of the random variables. (The functions $a + B\mathbf{x}$ and $A + B\mathbf{X}C$ are also called affine transformations, of $\mathbf{x}$ and $\mathbf{X}$, respectively.)

A vector $\mathbf{X}$ yields $p$ variances, the $\text{Var}[X_i]'s$, but also $(\binom{p}{2})$ covariances, the $\text{Cov}[X_i, X_j]'s$. These are usually conveniently arranged in a $p \times p$ matrix, the covariance matrix:

$$\Sigma = \text{Cov}(\mathbf{X}) = \begin{pmatrix}
\text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_p) \\
\text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_p) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \cdots & \text{Var}(X_p)
\end{pmatrix}. \hfill (3.98)

This matrix is symmetric, i.e., $\Sigma' = \Sigma$, where the prime means transpose. (The covariance matrix of a matrix $\mathbf{X}$ of random variables is typically defined by first changing the matrix $\mathbf{X}$ into a long vector, then defining the covariance of an affine transformation follows.

**Lemma 4.** For fixed $a$ and $B$,

$$\text{Cov}[a + B\mathbf{X}] = B \text{Cov}[\mathbf{X}]B'.$$  \hfill (3.101)

Note that this lemma is a matrix version of (3.45).

**Proof.**

$$\text{Cov}[a + B\mathbf{X}] = \text{Cov}[(a + B\mathbf{X} - E[a + B\mathbf{X}]) (a + B\mathbf{X} - E[a + B\mathbf{X}])'] \quad \text{by (3.99)}$$

$$= \text{Cov}[B(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])']$$

$$= B \text{Cov}[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])']B' \quad \text{by (3.97)}$$

$$= B \text{Cov}[\mathbf{X}]B' \quad \text{again by (3.99)}. \quad \square \hfill (3.102)$$

This lemma leads to a simple formula for the variance of $a + b_1X_1 + \cdots + b_pX_p$:

$$\text{Var}[a + b_1X_1 + \cdots + b_pX_p] = b'\text{Cov}[\mathbf{X}]b,$$  \hfill (3.103)

because we can write $a + b_1X_1 + \cdots + b_pX_p = a + b\mathbf{X}$ (so $a = a$ and $B = b'$ in (3.101).) Compare this formula to (3.47).
Marginal Distributions and Independence

4.1 Marginal distributions

Given the distribution of a vector of random variables, it is possible in principle to find the distribution of any individual component of the vector, or any subset of components. To illustrate, consider the distribution of the scores \((\text{Assignment}, \text{Exams})\) for a statistics class, where each variable has values “Lo” and “Hi”:

\[
\begin{array}{ccc}
\text{Assignments} & \downarrow & \text{Exams} \\
& \text{Lo} & \text{Hi} & \text{Sum} \\
\text{Lo} & 0.3365 & 0.1682 & 0.5047 \\
\text{Hi} & 0.1682 & 0.3271 & 0.4953 \\
\text{Sum} & 0.5047 & 0.4953 & 1
\end{array}
\] (4.1)

Thus about 33\% of the students did low on both assignments and exams, and about 33\% did high on both. But notice it is also easy to figure out the percentages of people who did low or high on the individual scores, e.g.,

\[
P[\text{Assignment} = \text{Lo}] = 0.5047 \text{ and (hence) } P[\text{Assignment} = \text{Hi}] = 0.4953. \] (4.2)

Coincidently, the exam scores have the same distribution. These numbers are in the margins of the table (4.1), hence the distribution of \(\text{Assignments}\) alone, and of \(\text{Exams}\) alone, are called marginal distributions. The distribution of \((\text{Assignments, Exams})\) together is called the joint distribution.

More generally, given the joint distribution of (the big vector) \((\mathbf{X}, \mathbf{Y})\), one can find the marginal distribution of the vector \(\mathbf{X}\) and the marginal distribution of the vector \(\mathbf{Y}\). (We don’t have to take consecutive components of the vector, e.g., given \((X_1, X_2, \ldots, X_5)\), we could be interested in the marginal distribution of \((X_1, X_3, X_4)\), say.)

Actually, the words joint and marginal can be dropped. The joint distribution of \((\mathbf{X}, \mathbf{Y})\) is just the distribution of \((\mathbf{X}, \mathbf{Y})\); the marginal distribution of \(\mathbf{X}\) is just the distribution of \(\mathbf{X}\) and the same for \(\mathbf{Y}\). These extra terms can be helpful, though, when dealing with different types of distributions in the same breath.
Before showing how to find the marginal distributions from the joint, we should deal with the spaces. Let \( W \) be the joint space of \((X, Y)\), and \( X' \) and \( Y' \) be the marginal spaces of \( X \) and \( Y \), respectively. Then

\[
X' = \{ x \mid (x, y) \in W \text{ for some } y \} \quad \text{and} \quad Y' = \{ y \mid (x, y) \in W \text{ for some } x \}.
\]

(4.3)

In Example 3.2.1, the joint space \( W \) is the upper left triangle of the unit square, hence the marginal spaces \( X' \) and \( Y' \) are both \((0, 1)\).

There are various approaches to finding the marginal distributions from the joint. First, suppose \( F(x, y) \) is the distribution function for \((X, Y)\) jointly, and \( F_X(x) \) is that for \( X \) marginally. Then (assuming \( X \) is \( p \times 1 \) and \( Y \) is \( q \times 1 \)),

\[
F_X(x) = P[X_1 \leq x_1, \ldots, X_p \leq x_p] = P[X_1 \leq x_1, \ldots, X_p \leq x_p, Y_1 \leq \infty, \ldots, Y_q \leq \infty] = F(x_1, \ldots, x_p, \infty, \ldots, \infty).
\]

(4.4)

That is, you put \( \infty \) in for the variables you are not interested in, because they are certainly less than infinity.

The mgf is equally easy. Suppose \( M(t, s) \) is the mgf for \((X, Y)\) jointly, so that

\[
M(t, s) = E[e^{t'X} + s'Y].
\]

(4.5)

To eliminate the dependence on \( Y \), we now set \( s \) to zero, that is, the mgf of \( X \) alone is

\[
M_X(t) = E[e^{t'X}] = E[e^{t'X} + 0'Y] = M(t, 0).
\]

(4.6)

### 4.1.1 Example: Marginals of a multinomial

Given \( X \sim \text{Multinomial}(n, p) \) as in (3.85), one may wish to find the marginal distribution of a single component, e.g., \( X_1 \). It should be Binomial, because now for each trial a success is that the observation is in the first category. To show this fact, we find the mgf of \( X_1 \) by setting \( t_2 = \cdots = t_K = 0 \) in (3.90):

\[
M_{X_1}(t) = M_X((t, 0, \ldots, 0)' = (p_1e^t + p_1 + \cdots + p_K)^n = (p_1e^t + 1 - p_1)^n,
\]

(4.7)

which is indeed the mgf of a Binomial as in (3.81). Specifically,

\[
X_1 \sim \text{Binomial}(n, p_1).
\]

(4.8)

### 4.2 Marginal densities

More challenging, but also more useful, is to find the marginal density from the joint density, assuming it exists. Suppose the joint distribution of the two random variables, \((X, Y)\), has pmf \( f(x, y) \), and space \( W \). Then \( X \) has a pmf, \( f_X(x) \), as well. To find it in terms of \( f \), write

\[
f_X(x) = P[X = x \text{ (and } Y \text{ can be anything})] = \sum_{y \mid (x, y) \in W} P[X = x, Y = y] = \sum_{y \mid (x, y) \in W} f(x, y)
\]

(4.9)
That is, you add up all the \( f(x,y) \) for that value of \( x \), as in the table (4.1). The same procedure works if \( X \) and \( Y \) are vectors. The set of \( y \)'s we are summing over we will call the \textbf{conditional space} of \( Y \) given \( X = x \), and denote \( \mathcal{Y}_x \):

\[
\mathcal{Y}_x = \{ y \in \mathcal{Y} \mid (x,y) \in \mathcal{W} \}. \tag{4.10}
\]

For Example 3.2.1, for any \( x \in (0,1) \), \( y \) ranges from \( x \) to 1, hence

\[
\mathcal{Y}_x = (x,1). \tag{4.11}
\]

In the coin Example 2.4.3, for any probability of heads \( x \), the range of \( Y \) is the same, so that \( \mathcal{Y}_x = \{0,1,\ldots,n\} \) for any \( x \in (0,1) \).

Thus in the general discrete case, we have

\[
f_X(x) = \sum_{y \in \mathcal{Y}_x} f(x,y), \quad x \in \mathcal{X}. \tag{4.12}
\]

### 4.2.1 Example: Ranks

The National Opinion Research Center Amalgam Survey of 1972 asked people to rank three types of areas in which to live: City over 50,000, Suburb (within 30 miles of a City), and Country (everywhere else). The table below shows the results\(^1\), with respondents categorized by their current residence.

<table>
<thead>
<tr>
<th>Ranking (City, Suburb, Country)</th>
<th>Residence City</th>
<th>Residence Suburb</th>
<th>Residence Country</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 2, 3)</td>
<td>210</td>
<td>22</td>
<td>10</td>
<td>242</td>
</tr>
<tr>
<td>(1, 3, 2)</td>
<td>23</td>
<td>4</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>(2, 1, 3)</td>
<td>111</td>
<td>45</td>
<td>14</td>
<td>170</td>
</tr>
<tr>
<td>(2, 3, 1)</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>(3, 1, 2)</td>
<td>204</td>
<td>299</td>
<td>125</td>
<td>628</td>
</tr>
<tr>
<td>(3, 2, 1)</td>
<td>81</td>
<td>126</td>
<td>152</td>
<td>359</td>
</tr>
<tr>
<td>Total</td>
<td>637</td>
<td>500</td>
<td>302</td>
<td>1439</td>
</tr>
</tbody>
</table>

That is, a ranking of \( (1,2,3) \) means that person ranks living in the city best, suburbs next, and country last. There were 242 people in the sample with that ranking, 210 of whom live in the city (so they should be happy), 22 of whom live in the suburbs, and just 10 of whom live in the country.

The random vector here is \((X,Y,Z)\), say, where \( X \) represents the rank of City, \( Y \) that of Suburb, and \( Z \) that of Country. The space consists of the six permutations of 1,2 and 3:

\[
\mathcal{W} = \{(1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1)\}, \tag{4.14}
\]

as in the first column of the table. Suppose the total column is our population, so that there are 1439 people all together, and we randomly choose a person from this

---

population. Then the (joint) distribution of the person’s ranking \((X, Y, Z)\) is given by

\[
f(x, y, z) = P[(X, Y, Z) = (x, y, z)] = \begin{cases} 
242/1439 & \text{if } (x, y, z) = (1, 2, 3) \\
28/1439 & \text{if } (x, y, z) = (1, 3, 2) \\
170/1439 & \text{if } (x, y, z) = (2, 1, 3) \\
12/1439 & \text{if } (x, y, z) = (2, 3, 1) \\
628/1439 & \text{if } (x, y, z) = (3, 1, 2) \\
359/1439 & \text{if } (x, y, z) = (3, 2, 1) 
\end{cases} \tag{4.15}
\]

This distribution could use with some summarizing, e.g., what are the marginal distributions of \(X, Y,\) and \(Z\)? For each ranking \(x = 1, 2, 3,\) we have to add over the possible rankings of \(Y\) and \(Z,\) so that

\[
f_X(1) = f(1, 2, 3) + f(1, 3, 2) = \frac{28 + 242}{1439} = 0.1876; \\
f_X(2) = f(2, 1, 3) + f(2, 3, 1) = \frac{170 + 12}{1439} = 0.1265; \\
f_X(3) = f(3, 1, 2) + f(3, 2, 1) = \frac{628 + 359}{1439} = 0.6859. \tag{4.16}
\]

Thus City is ranked third over 2/3 of the time. The marginal rankings of Suburb and Country can be obtained similarly. \(\square\)

Again for two variables, suppose now that the pdf is \(f(x, y).\) We know that the distribution function is related to the pdf via

\[
F(x, y) = \int_{(-\infty, x) \cap \mathcal{X}} \int_{(-\infty, y) \cap \mathcal{Y}} f(u, v) dv du. \tag{4.17}
\]

From (4.4), to obtain the distribution function, we set \(y = \infty,\) which means in the inside integral, we can remove the "\((-\infty, y)"\) part:

\[
F_X(x) = \int_{(-\infty, x) \cap \mathcal{X}} \int_{\mathcal{Y}_x} f(u, v) dv du. \tag{4.18}
\]

Then the pdf of \(X\) is found by taking the derivative with respect to \(x\) for \(x \in \mathcal{X},\) which here just means stripping away the outer integral and setting \(u = x\) (and \(v = y,\) if we wish):

\[
f_X(x) = \frac{\partial}{\partial x} F_X(x) = \int_{\mathcal{Y}_x} f(x, y) dy. \tag{4.19}
\]

Thus instead of summing over the \(y\) as in (4.12), we integrate. This procedure is often called "integrating out \(y)."

Turning to Example 3.2.1 again, we have \(f(x, y) = 2\) for \((x, y) \in \mathcal{W}.\) From (4.11), for \(x \in (0, 1),

\[
f_X(x) = \int_{\mathcal{Y}_x} f(x, y) dy = \int_x^1 2 dy = 2(1 - x). \tag{4.20}
\]

With vectors, the process is the same, just add lines under the symbols:

\[
f_X(x) = \int_{\mathcal{Y}_x} f(x, y) dy. \tag{4.21}
\]
4.3 Independence

Much of statistics is geared towards evaluation of relationships between variables: Does smoking cause cancer? Do cell phones? What factors explain the rise in asthma? The absence of a relationship, independence, is also important.

**Definition 2.** Suppose \((X, Y)\) has joint distribution \(P\), and marginal spaces \(X\) and \(Y\), respectively. Then \(X\) and \(Y\) are independent if

\[
P[X \in A \text{ and } Y \in B] = P[X \in A] \times P[Y \in B] \quad \text{for all } A \subset X \text{ and } B \subset Y. \tag{4.22}
\]

Also, if \((X^{(1)}, \ldots, X^{(K)})\) has distribution \(P\), and the vector \(X^{(k)}\) has space \(X_k\), then \(X^{(1)}, \ldots, X^{(K)}\) are (mutually) independent if

\[
P[X^{(1)} \in A_1, \ldots, X^{(K)} \in A_K] = P[X^{(1)} \in A_1] \times \cdots \times P[X^{(K)} \in A_K] \quad \text{for all } A_1 \subset X_1, \ldots, A_K \subset X_K. \tag{4.23}
\]

The basic idea in independence is that what happens with one variable does not affect what happens with another. There are a number of useful equivalences for independence of \(X\) and \(Y\). (Those for mutual independence of \(K\) vectors hold similarly.)

- **Distribution functions.** \(X\) and \(Y\) are independent if and only if
  \[
  F(x, y) = F_X(x) \times F_Y(y) \quad \text{for all } x \in \mathbb{R}^p, y \in \mathbb{R}^q; \tag{4.24}
  \]

- **Products of functions.** \(X\) and \(Y\) are independent if and only if
  \[
  E[g(X)h(Y)] = E[g(X)] \times E[h(Y)] \tag{4.25}
  \]
  for all functions \(g : X \rightarrow \mathbb{R}\) and \(h : Y \rightarrow \mathbb{R}\) whose means exist;

- **MGF’s.** Suppose the marginal mgf’s of \(X\) and \(Y\) are finite for \(t\) and \(s\) in neighborhoods of zero (respectively in \(\mathbb{R}^p\) and \(\mathbb{R}^q\)). Then \(X\) and \(Y\) are independent if and only if
  \[
  M(L, s) = M_X(t)M_Y(s) \tag{4.26}
  \]
  for all \((L, s)\) in a neighborhood of zero in \(\mathbb{R}^{p+q}\).

**Remark.** The second item can be used to show that independent random variables are uncorrelated, because as in (16.123), \(\text{Cov}[X, Y] = E[XY] - E[X]E[Y]\), and (4.25) shows that \(E[XY] = E[X]E[Y]\) if \(X\) and \(Y\) are independent. Be aware that the implication does not go the other way, that is, \(X\) and \(Y\) can have correlation 0 and still not be independent. For example, suppose \(W = \{(0,1), (0,-1), (1,0), (-1,0)\}\), and \(P[(X, Y) = (x, y)] = \frac{1}{4}\) for each \((x, y) \in W\). Then it is not hard to show that \(E[X] = E[Y] = 0\), and that \(E[XY] = 0\) (in fact, \(XY = 0\) always), hence \(\text{Cov}[X, Y] = 0\). But \(X\) and \(Y\) are not independent, e.g., take \(A = \{0\}\) and \(B = \{0\}\). Then

\[
P[X \in A \text{ and } Y \in B] = P[X = 0 \text{ and } Y = 0] = 0 \neq P[X = 0]P[Y = 0] = \frac{1}{2} \times \frac{1}{2}. \quad \Box \tag{4.27}
\]
4.3.1 Example: Independent exponentials

Suppose $U$ and $V$ are independent Exponential(1)’s. The mgf of an Exponential(1) is $1/(1-t)$ for $t < 1$. See (3.74), which gives the mgf of Gamma($\alpha$, $\lambda$) as $(\lambda/(\lambda - t))^\alpha$ for $t < \lambda$. Thus the mgf of $(U, V)$ is

$$M_{(U,V)}(t_1, t_2) = M_U(t_1)M_V(t_2) = \frac{1}{1-t_1} \frac{1}{1-t_2}, \quad t_1 < 1, t_2 < 1.$$  \hfill (4.28)

Now let

$$X = U + V \quad \text{and} \quad Y = X - Y.$$  \hfill (4.29)

Are $X$ and $Y$ independent? What are their marginal distributions? We can start by looking at the mgf:

$$M_{(X,Y)}(s_1, s_2) = E[e^{s_1X+s_2Y}] = E[e^{s_1(U+V)+s_2(U-V)}] = E[e^{s_1+s_2}U+(s_1-s_2)V] = M_{(U,V)}(s_1 + s_2, s_1 - s_2) = \frac{1}{1-s_1-s_2} \frac{1}{1-s_1+s_2}. \hfill (4.30)$$

This mgf is finite if $s_1 + s_2 < 1$ and $s_1 - s_2 < 1$, which is a neighborhood of $(0, 0)$. If $X$ and $Y$ are independent, then this mgf must factor into the two individual mgf’s. It does not appear to factor. More formally, the marginal mgf’s are

$$M_X(s_1) = M_{(X,Y)}(s_1, 0) = \frac{1}{(1-s_1)^2}, \hfill (4.31)$$

and

$$M_Y(s_2) = M_{(X,Y)}(0, s_2) = \frac{1}{(1-s_2)(1+s_2)} = \frac{1}{1-s_2^2}. \hfill (4.32)$$

Note that the first one is that of a Gamma(2, 1), so that $X \sim \text{Gamma}(2, 1)$. The one for $Y$ may not be recognizable, but it turns out to be the mgf of a double exponential. But,

$$M_{(X,Y)}(s_1, s_2) \neq M_X(s_1)M_Y(s_2), \hfill (4.33)$$

hence $X$ and $Y$ are not independent. They are uncorrelated, however:

$$\text{Cov}[X,Y] = \text{Cov}[U+V, U-V] = \text{Var}[U] - \text{Var}[V] - \text{Cov}[U,V] + \text{Cov}[V,U] = 1 - 1 = 0.$$  \hfill (4.34)

4.3.2 Spaces and densities

Suppose $(X,Y)$ is discrete, with pmf $f(x,y) > 0$ for $(x,y) \in \mathcal{W}$, and $f_X$ and $f_Y$ are the marginal pmf’s of $X$ and $Y$, respectively. Then applying (4.22) to the singleton sets $\{x\}$ and $\{y\}$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ shows that

$$P[X = x, Y = y] = P[X = x] \times P[Y = y], \hfill (4.35)$$
which translates to
\[ f(x, y) = f_X(x)f_Y(y). \tag{4.36} \]
In particular, this equation shows that if \( x \in \mathcal{X} \) and \( y \in \mathcal{Y} \), then \( f(x, y) > 0 \), hence \((x, y) \in W\). That is, if \( X \) and \( Y \) are independent, then
\[ W = \mathcal{X} \times \mathcal{Y}, \tag{4.37} \]
the “rectangle” created from the marginal spaces.

Formally, given sets \( A \subset \mathbb{R}^p \) and \( B \subset \mathbb{R}^q \), the \textbf{rectangle} \( A \times B \) is defined to be
\[ A \times B = \{(x, y) \mid x \in A \text{ and } y \in B\} \subset \mathbb{R}^{p+q}. \tag{4.38} \]
The set may not be a rectangle in the usual sense, although it will be if \( p = q = 1 \) and \( A \) and \( B \) are both intervals. Figure 4.1 has some examples. Of course, \( \mathbb{R}^{p+q} \) is a rectangle itself, being \( \mathbb{R}^p \times \mathbb{R}^q \). The result (4.37) holds in general.

**Lemma 5.** If \( X \) and \( Y \) are independent, then the spaces can be taken so that (4.37) holds.

This lemma implies that if the joint space is not a rectangle, then \( X \) and \( Y \) are not independent. Consider example 3.2.1, where \( W = \{(x, y) \mid 0 < x < y < 1\} \), a triangle. If we take a square below that triangle, such as \((0.8, 1) \times (0.8, 1)\), then
\[ P[(X, Y) \in (0.8, 1) \times (0.8, 1)] = 0 \quad \text{but} \quad P[X \in (0.8, 1)]P[Y \in (0.8, 1)] > 0, \tag{4.39} \]
so that \( X \) and \( Y \) are not independent. The result extends to more variables. Consider Example 4.2.1 on ranks. Here, the marginal spaces are
\[ \mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{1, 2, 3\}, \tag{4.40} \]
but
\[ W \neq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} = \{(1, 1, 1), (1, 1, 2), (1, 1, 3), \ldots, (3, 3, 3)\}, \tag{4.41} \]
in particular because \( W \) has only 6 elements, while the product space has \( 3^3 = 27 \).

The factorization in (4.35) is necessary and sufficient for independence in the discrete and continuous cases.

**Lemma 6.** Suppose \( X \) has marginal density \( f_X(x) \), and \( Y \) has marginal density \( f_Y(y) \), and they are both pdf’s or both pmf’s. Then \( X \) and \( Y \) are independent if and only if the distribution of \((X, Y)\) is given by density
\[ f(x, y) = f_X(x)f_Y(y), \tag{4.42} \]
and space
\[ W = \mathcal{X} \times \mathcal{Y}. \tag{4.43} \]

We can simplify a little, that is, as long as the space and joint pdf factor, we have independence.

**Lemma 7.** Suppose \((X, Y)\) has joint density \( f(x, y) \). Then \( X \) and \( Y \) are independent if and only if the density can be written as
\[ f(x, y) = g(x)h(y) \tag{4.44} \]
for some functions \( g \) and \( h \), and
\[ W = \mathcal{X} \times \mathcal{Y}. \tag{4.45} \]

This lemma is not presuming the \( g \) and \( h \) are actual densities, although they certainly could be.
Figure 4.1: Some rectangles
4.3.3 IID

A special case of independence has the vectors with the exact same distribution, as well as being independent, that is, \( \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)} \) are independent, and all have the same marginal distribution. We say the vectors are iid, meaning "independent and identically distributed." This type of distribution often is used to model random samples, where \( n \) individuals are chosen from a (virtually) infinite population, and \( p \) variables are recorded on each. Then \( K = n \) and the \( \mathbf{X}^{(i)} \)'s are \( p \times 1 \) vectors. If the marginal density is \( f_{\mathbf{X}^{(i)}} \) for each \( \mathbf{X}^{(i)} \), with marginal space \( \mathcal{X}_0 \), then the joint density of the entire sample is

\[
f(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}) = f_{\mathbf{X}^{(1)}}(\mathbf{x}^{(1)}) \cdots f_{\mathbf{X}^{(n)}}(\mathbf{x}^{(n)})
\]

with space

\[
\mathcal{X} = \mathcal{X}_0 \times \cdots \times \mathcal{X}_0 = (\mathcal{X}_0)^n.
\]
Chapter 5

Transformations I

A major task of mathematical statistics is finding, or approximating, the distributions of random variables that are functions of other random variables. For example, given \( X_1, \ldots, X_n \) iid observations, find the distribution (exactly or approximately) of the sample mean,

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.
\]  

(5.1)

This chapter will address finding the exact distribution. There are many approaches, and which one to use may not always be obvious. The following sections run through a number of possibilities.

5.1 Distribution functions

Suppose the \( p \times 1 \) vector \( \mathbf{X} \) has distribution function \( F_X \), and \( Y = g(\mathbf{X}) \) for some function

\[
g : \mathcal{X} \rightarrow \mathcal{Y} \subset \mathbb{R}.
\]  

(5.2)

Then the distribution function \( F_Y \) of \( Y \) is

\[
F_Y(y) = P[Y \leq y] = P[g(\mathbf{X}) \leq y], \ y \in \mathbb{R}.
\]  

(5.3)

The final probability is in principle obtainable from the distribution of \( \mathbf{X} \), which solves the problem. If \( Y \) has a pdf, we can then find that by differentiating.

5.1.1 Example: Sum of uniforms

Suppose the \( p \times 1 \) vector \( \mathbf{X} \) has distribution function \( F_X \), and \( Y = g(\mathbf{X}) \) for some function

\[
g : \mathcal{X} \rightarrow \mathcal{Y} \subset \mathbb{R}.
\]  

(5.2)

Then the distribution function \( F_Y \) of \( Y \) is

\[
F_Y(y) = P[Y \leq y] = P[g(\mathbf{X}) \leq y], \ y \in \mathbb{R}.
\]  

(5.3)

The final probability is in principle obtainable from the distribution of \( \mathbf{X} \), which solves the problem. If \( Y \) has a pdf, we can then find that by differentiating.

As in (5.2), its distribution function is

\[
F_Y(y) = P[Y \leq y] = P[X_1 + X_2 \leq y].
\]  

(5.5)
Note that it is important in these calculations to distinguish between random variables (the capitalized letters, $X_1, X_2, Y$), and particular values they could take on (the lower case letters, like $y$).

The space of $Y$ is $\mathcal{Y} = (0, 2)$. Because the pdf of $X$ is 1, the probability of any set $A \subset \mathcal{X}$ is just its area. Figure 5.1 shows the regions $\{ \mathbf{x} \mid x_1 + x_2 \leq y \}$ for $y = 0.8$ and $y = 1.3$. When $0 < y < 1$, this region is a triangle, whose area is easily seen to be $y^2/2$. When $1 \leq y < 2$, the area is the unit square minus a triangle, $1 - (2 - y)^2/2$. Thus

$$F_Y(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ y^2/2 & \text{if } 0 < y < 1 \\ 1 - (2 - y)^2/2 & \text{if } 1 \leq y < 2 \\ 1 & \text{if } 2 \leq y \end{cases}. \quad (5.6)$$

The pdf is then the derivative, so that for $0 < y < 2$,

$$f_Y(y) = \begin{cases} y & \text{if } 0 < y < 1 \\ 2 - y & \text{if } 1 \leq y < 2 \end{cases}. \quad (5.7)$$

which is a “tent” distribution, as in Figure 5.2.
5.1. Distribution functions

5.1.2 Example: Chi-squared on 1 df

If $X \sim N(\mu, 1)$, then $Y = X^2$ is **noncentral chisquared** on 1 degree of freedom with **noncentrality parameter** $\Delta = \mu^2$, written $Y \sim \chi_1^2(\Delta)$. (5.8)

(In Definition 10, we present the more general chi-squared.) Because $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = (0, \infty)$. (Actually, it is $[0, \infty)$, but leaving out one point is ok for continuous distributions.) Letting $f_X(x \mid \mu)$ be the pdf of $X$, the distribution function of $Y$ is, for $y > 0$,

$$
F_Y(y \mid \mu) = P[X^2 \leq y] = \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x \mid \mu) dx. \quad (5.9)
$$

The derivative will then yield the pdf of $Y$. We need the chain rule result from calculus:

$$
\frac{\partial}{\partial y} \int_{a(y)}^{b(y)} g(x) dx = g(b(y))b'(y) - g(a(y))a'(y). \quad (5.10)
$$

Applied to (5.9) gives

$$
f_Y(y \mid \mu) = F_Y'(y) = f_X(\sqrt{y} \mid \mu) \frac{\partial \sqrt{y}}{\partial y} - f_X(-\sqrt{y} \mid \mu) \frac{\partial (-\sqrt{y})}{\partial y} = \frac{1}{2\sqrt{y}} (f_X(\sqrt{y} \mid \mu) + f_X(-\sqrt{y} \mid \mu)). \quad (5.11)
$$
Note that this equation is not restricted to $f_X$ being normal. Now the $N(\mu, 1)$ density is

$$f_X(x \mid \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (x-\mu)^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} x^2 e^{-\frac{1}{2} \mu^2} e^{\mu x}}, \quad (5.12)$$

so that we can write

$$f_Y(y \mid \mu) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2} y} \left( \frac{e^{\mu \sqrt{y}} + e^{-\mu \sqrt{y}}}{2} \right) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2} y} e^{-\frac{1}{2} \mu^2} \cosh(\mu \sqrt{y}). \quad (5.13)$$

(The $\cosh$ is the hyperbolic cosine.) The notation in (5.8) implies that the distribution of $Y$ depends on $\mu$ through only $\Delta = \mu^2$, that is, the distribution is the same for $\mu$ and $-\mu$. In fact, that can be seen to be true because $f_Y(y \mid \mu) = f_Y(y \mid -\mu)$, so we can replace $\mu$ with $\sqrt{\Delta}$ in (5.13).

When $\Delta = 0$ (i.e., $\mu = 0$), $Y$ is then central chi-squared, written

$$Y \sim \chi^2_1. \quad (5.14)$$

It’s pdf is (5.13) without the last two expressions:

$$f_Y(y \mid 0) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2} y}. \quad (5.15)$$

Notice that this pdf is that of $\text{Gamma}(\frac{1}{2}, \frac{1}{2})$. (Also, we can see that $\Gamma(1/2) = \sqrt{\pi}$, as mentioned in (3.70)).

### 5.1.3 Example: Probability transform

Generating random numbers on a computer is a common activity in statistics, e.g., for inference, to randomize subjects to treatments, and to assess the performance of various procedures. It is easy to generate independent $\text{Uniform}(0, 1)$ random variables — Easy in the sense that many people have worked very hard for many years to develop good methods that are now available in most statistical software. Actually, the numbers are not truly random, but rather pseudo-random, because there is a deterministic algorithm producing them. But you could also question whether randomness exists in the real world anyway. Even flipping a coin is deterministic, if you know all the physics in the flip.

One usually is not satisfied with uniforms, but rather has normals or gammas, or something more complex, in mind. There are many clever ways to create the desired random variables from uniforms (see for example the Box-Mueller transformation in Equation 6.82), but the most basic uses the inverse distribution function. We suppose $U \sim \text{Uniform}(0, 1)$, and wish to generate a $X$ that has given distribution function $F$. Assume that $F$ is continuous, and strictly increasing for $x \in X$, so that $F^{-1}(u)$ is well-defined for each $u \in (0, 1)$. Consider the random variable

$$W = F^{-1}(U). \quad (5.16)$$

Then its distribution function is

$$F_W(w) = P[W \leq w] = P[F^{-1}(U) \leq w] = P[U \leq F(w)] = F(w), \quad (5.17)$$
where the last step follows because \( U \sim \text{Uniform}(0,1) \) and \( 0 \leq F(w) \leq 1 \). But that equation means that \( W \) has the desired distribution function, hence to generate an \( X \), we generate a \( U \) and take \( X = F^{-1}(U) \).

To illustrate, suppose we wish \( X \sim \text{Exponential}(\lambda) \). The distribution function \( F \) of \( X \) is zero for \( x \leq 0 \), and
\[
F(x) = \int_0^x \lambda e^{-\lambda w} \, dw = 1 - e^{-\lambda x} \quad \text{for} \quad x > 0. \tag{5.18}
\]
Thus for \( u \in (0,1) \),
\[
u = F(x) \implies x = F^{-1}(u) = -\log(1-u)/\lambda. \tag{5.19}
\]

One limitation of this method is that \( F^{-1} \) is not always computationally simple. Even in the normal case there is no closed form expression. In addition, this method does not work directly for generating multivariate \( X \), because then \( F \) is not invertible.

The approach works for non-continuous distributions as well, but care must be taken because \( u \) may fall in a gap. For example, suppose \( X \) is Bernoulli(\( \frac{1}{2} \)). Then for \( u = \frac{1}{2} \), we have \( x = 0 \), since \( F(0) = \frac{1}{2} \), but no other value of \( u \in (0,1) \) has an \( x \) with \( F(x) = u \). The fix is for \( 0 < u < \frac{1}{2} \), to set \( F^{-1}(u) = 0 \), and for \( \frac{1}{2} < u < 1 \), to set \( F^{-1}(u) = 1 \), so that half the time \( F^{-1}(U) \) is 0, and half the time it is 1. That is, if \( u \) is in a gap, set \( F^{-1}(u) \) to the value of \( x \) for that gap. Mathematically, we define for \( 0 < u < 1 \),
\[
F^{-1}(u) = \min\{x \mid F(x) \geq u\}, \tag{5.20}
\]
which will work for continuous and noncontinuous \( F \).

The process can be reversed, which is useful in hypothesis testing, to obtain \( p \)-values. That is, suppose \( X \) has continuous strictly increasing (on \( X \)) distribution function \( F \), and let
\[
U = F(X). \tag{5.21}
\]
Then the distribution function of \( U \) is
\[
F_U(u) = P[F(X) \leq u] = P[X \leq F^{-1}(u)] = F(F^{-1}(u)) = u, \tag{5.22}
\]
which is the distribution function of a \( \text{Uniform}(0,1) \), i.e.,
\[
F(X) \sim \text{Uniform}(0,1). \tag{5.23}
\]

### 5.2 Moment generating functions

Instead of finding the distribution function of \( Y \), one could try to find the mgf. By the uniqueness Theorem 1, if you recognize the mgf of \( Y \) as being for a particular distribution, then you know \( Y \) has that distribution.
5.2.1 Example: Uniform → Exponential

For example, suppose \( X \sim \text{Uniform}(0, 1) \), and \( Y = -\log(X) \). Then the mgf of \( Y \) is

\[
M_Y(t) = E[e^{tY}] = E[e^{-t\log(X)}] = \int_0^1 e^{-t\log(x)} \, dx = \int_0^1 x^{-t} \, dx = \frac{1}{-t+1} x^{-t+1} \bigg|_{x=0}^1 = \frac{1}{1-t} \quad \text{if } t < 1 \quad \text{(and } +\infty \text{ if not).} \tag{5.24}
\]

It turns out that this mgf is that of the Gamma in (3.74), where \( \alpha = \lambda = 1 \), which means it is also Exponential(1), as in (3.72). Thus the pdf of \( Y \) is

\[
f_Y(y) = e^{-y} \quad \text{for } y \in \mathcal{Y} = (0, \infty). \tag{5.25}
\]

5.2.2 Example: Sums of independent Gamma’s

The mgf approach is especially useful for the sums of independent random variables. For example, suppose \( X_1, \ldots, X_K \) are independent, with \( X_k \sim \text{Gamma}(\alpha_k, \lambda) \). (They all have the same scale, but are allowed different shapes.) Let \( Y = X_1 + \cdots + X_K \).

Then its mgf is

\[
M_Y(t) = E[e^{tY}] = E[e^{t(X_1+\cdots+X_K)}] = E[e^{tX_1} \cdots e^{tX_K}] \quad \text{by independence of the } X_k \text{'s} = M_1(t) \times \cdots \times M_K(t) \quad \text{where } M_k(t) \text{ is the mgf of } X_k
\]

\[
= \left( \frac{\lambda}{\lambda - t} \right)^{\alpha_1} \cdots \left( \frac{\lambda}{\lambda - t} \right)^{\alpha_K} \quad \text{for } t < \lambda, \text{ by (3.74)}
\]

\[
= \left( \frac{\lambda}{\lambda - t} \right)^{\alpha_1+\cdots+\alpha_K}. \tag{5.26}
\]

But then this mgf is that of \( \text{Gamma}(\alpha_1 + \cdots + \alpha_K, \lambda) \). Thus

\[
X_1 + \cdots + X_K \sim \text{Gamma}(\alpha_1 + \cdots + \alpha_K, \lambda). \tag{5.27}
\]

What if the \( \lambda \)’s are not equal? Then it is still easy to find the mgf, but it would not be the Gamma mgf, or anything else we have seem so far.

5.2.3 Linear combinations of independent Normals

In (3.65), we have that the mgf of a \( N(\mu, \sigma^2) \) is

\[
M(t \mid \mu, \sigma^2) = e^{t\mu + \frac{1}{2} t^2 \sigma^2}, \quad t \in \mathbb{R}. \tag{5.28}
\]
Now suppose \( X_1, \ldots, X_K \) are independent, \( X_k \sum N(\mu_k, \sigma_k^2) \). Consider the affine combination
\[
Y = a + b_1 X_1 + \cdots + b_K X_K.
\] (5.29)

It is straightforward, from Section 3.2.2, to see that
\[
E[Y] = a + b_1 \mu_1 + \cdots + b_K \mu_K \quad \text{and} \quad \text{Var}[Y] = b_1^2 \sigma_1^2 + \cdots + b_K^2 \sigma_K^2,
\] (5.30)
since independence implies that all the covariances are 0. But those equations do not give the entire distribution of \( Y \). We need the mgf:
\[
M_Y(t) = E[e^{iY}] = E[e^{i(t(a + b_1 X_1 + \cdots + b_K X_K))}]
\]
\[
\quad = e^{iaX} E[e^{i(ta + b_1 X_1 + \cdots + b_K X_K)}]
\]
\[
\quad = e^{iaX} M(ta | \mu_1, \sigma_1^2 \times \cdots \times M(ta | \mu_K, \sigma_K^2)
\]
\[
\quad = e^{iaX} e^{tb_1 \mu_1 + \frac{t^2}{2} b_1^2 \sigma_1^2} \cdots \times e^{tb_k \mu_k + \frac{t^2}{2} b_k^2 \sigma_k^2}
\]
\[
\quad = M(t | a + b_1 \mu_1 + \cdots + b_K \mu_K, b_1^2 \sigma_1^2 + \cdots + b_K^2 \sigma_K^2). \quad (5.31)
\]

Notice that in going from the third to fourth step, we have changed the variable for the mgf’s from \( t \) to the \( b_k t \)’s, which is legitimate. That mgf is indeed the mgf of a Normal, with the appropriate mean and variance, i.e.,
\[
Y \sim N(a + b_1 \mu_1 + \cdots + b_K \mu_K, b_1^2 \sigma_1^2 + \cdots + b_K^2 \sigma_K^2). \quad (5.32)
\]

If the \( X_k \)'s are iid \( N(\mu, \sigma^2) \), then (5.32) can be used to show that
\[
\sum X_k \sim N(K\mu, K\sigma^2) \quad \text{and} \quad \overline{X} \sim N(\mu, \sigma^2 / K). \quad (5.33)
\]

### 5.2.4 Normalized means

The Central Limit Theorem is central to statistics as it justifies using the normal distribution in certain non-normal situations. Here we assume we have \( X_1, \ldots, X_n \) iid with any distribution, as long as the mgf \( M(t) \) of \( X_i \) exists for \( t \) in a neighborhood of 0. The normalized mean is
\[
W_n = \sqrt{n} \frac{\overline{X} - \mu}{\sigma}, \quad (5.34)
\]
where \( \mu = E[X_i] \) and \( \sigma^2 = \text{Var}[X_i] \). If the \( X_i \) are normal, then \( W_n \sim N(0, 1) \). The Central Limit Theorem implies that even if the \( X_i \)'s are not normal, \( W_n \) is “approximately” normal if \( n \) is “large.” We will look at the cumulant generating function of \( W_n \), and compare it to the normal’s.

First, the mgf of \( W_n \):
\[
M_n(t) = E[e^{iW_n}] = E[e^{i\sqrt{n}(\overline{X} - \mu) / \sigma}]
\]
\[
\quad = e^{-i\sqrt{n} \mu / \sigma} E[e^{it / (\sqrt{n}\sigma)} \sum x_i]
\]
\[
\quad = e^{-i\sqrt{n} \mu / \sigma} \prod E[e^{it / (\sqrt{n}\sigma)} x_i]
\]
\[
\quad = e^{-i\sqrt{n} \mu / \sigma} M(t / (\sqrt{n}\sigma))^n. \quad (5.35)
\]
Letting $c(t)$ be the cumulant generating function of $X_i$, and $c_n(t)$ be that of $W_n$, we have that

$$c_n(t) = -\frac{t\sqrt{n\mu}}{\sigma} + nc\left(\frac{t}{\sqrt{n\sigma}}\right).$$  \hspace{1cm} (5.36)

Now imagine taking the derivatives of $c_n(t)$ with respect to $t$. The first two:

$$c_n'(t) = -\frac{\sqrt{n\mu}}{\sigma} + nc'\left(\frac{t}{\sqrt{n\sigma}}\right) \Rightarrow E[W_n] = c_n'(0) = -\frac{\sqrt{n\mu}}{\sigma} + n\mu \frac{1}{\sqrt{n\sigma}} = 0;$$

$$c_n''(t) = nc''\left(\frac{t}{\sqrt{n\sigma}}\right) \left(\frac{1}{\sqrt{n\sigma}}\right)^2 \Rightarrow \text{Var}[W_n] = c_n''(0) = n\sigma^2 \frac{1}{(\sqrt{n\sigma})^2} = 1.$$  \hspace{1cm} (5.37)

It is easy enough to find the mean and variance of $W_n$ directly, but this approach is helpful for the higher-order cumulants. Note that each derivative brings out another $1/\sqrt{n\sigma}$ from $c(t)$, hence the $k^{th}$ cumulant of $W_n$ is

$$c_n^{(k)}(0) = nc^{(k)}(0) \frac{1}{(\sqrt{n\sigma})^k},$$  \hspace{1cm} (5.38)

and letting $\gamma_k$ be the $k^{th}$ cumulant of $X_i$, and $\gamma_{n,k}$ be that of $W_n$,

$$\gamma_{n,k} = \frac{1}{n^{k/2-1}} \frac{1}{\sigma^k} \gamma_k.$$  \hspace{1cm} (5.39)

For the normal, all cumulants are 0 after the first two. Thus the closer the $\gamma_{n,k}$’s are to 0, the closer the distribution of $W_n$ is to $N(0, 1)$ in some sense. There are then two factors: The larger $n$, the closer these cumulants are to 0 for $k > 2$. Also, the smaller $\gamma_k/\sigma^k$ in absolute value, the closer to 0.

For example, if the $X_i$’s are Exponential(1), or Gamma(1, 1), then from (3.77) the cumulants are $\gamma_k = (k-1)!$, and $c_{n,k} = (k-1)!/n^{k/2-1}$. The Poisson($\lambda$) has cumulant generating function $c(t) = \lambda(e^t - 1)$, so all its derivatives are $\lambda e^t$, hence $\gamma_k = \lambda$, and

$$\gamma_{n,k} = \frac{1}{n^{k/2-1}} \frac{1}{\sigma^{k/2}} \gamma_k = \frac{1}{n^{k/2-1}} \frac{1}{\lambda^{k/2}} \lambda = \frac{1}{(n\lambda)^{k/2-1}}.$$  \hspace{1cm} (5.40)

The DoubleExponential has pdf $(1/2)e^{-|x|}$ for $x \in \mathbb{R}$. It’s cumulants are

$$\gamma_k = \begin{cases} 
0 & \text{if } k \text{ is odd} \\
2(k-1)! & \text{if } k \text{ is even}.
\end{cases}$$  \hspace{1cm} (5.41)

Thus the variance is 2, and

$$\gamma_{n,k} = \gamma_k = \begin{cases} 
0 & \text{if } k \text{ is odd} \\
(k-1)!/(2n)^{k/2-1} & \text{if } k \text{ is even}.
\end{cases}$$  \hspace{1cm} (5.42)
Here is a small table with some values of $\gamma_{k,n}$:

<table>
<thead>
<tr>
<th>$\gamma_{n,k}$</th>
<th>Normal</th>
<th>$Exp(1)$</th>
<th>Pois$(1/10)$</th>
<th>Pois$(1)$</th>
<th>Pois$(10)$</th>
<th>DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 3$ $n = 1$</td>
<td>0</td>
<td>2.000</td>
<td>3.162</td>
<td>1.000</td>
<td>0.316</td>
<td>0.000</td>
</tr>
<tr>
<td>$n = 10$</td>
<td>0</td>
<td>0.632</td>
<td>1.000</td>
<td>0.316</td>
<td>0.100</td>
<td>0.000</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>0</td>
<td>0.200</td>
<td>0.316</td>
<td>0.100</td>
<td>0.032</td>
<td>0.000</td>
</tr>
<tr>
<td>$k = 4$ $n = 1$</td>
<td>0</td>
<td>6.000</td>
<td>10.000</td>
<td>1.000</td>
<td>0.100</td>
<td>3.000</td>
</tr>
<tr>
<td>$n = 10$</td>
<td>0</td>
<td>0.600</td>
<td>1.000</td>
<td>0.100</td>
<td>0.010</td>
<td>0.000</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>0</td>
<td>0.060</td>
<td>0.100</td>
<td>0.010</td>
<td>0.001</td>
<td>0.030</td>
</tr>
<tr>
<td>$k = 5$ $n = 1$</td>
<td>0</td>
<td>24.000</td>
<td>31.623</td>
<td>1.000</td>
<td>0.032</td>
<td>0.000</td>
</tr>
<tr>
<td>$n = 10$</td>
<td>0</td>
<td>0.759</td>
<td>1.000</td>
<td>0.032</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>0</td>
<td>0.024</td>
<td>0.032</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

For each distribution, as $n$ increases, the cumulants do decrease. Also, the exponential is closer to normal than the Poisson$(1/10)$, but the Poisson$(1)$ is closer than the exponential, and the Poisson$(10)$ is even closer. The Double Exponential is symmetric, so its odd cumulants are 0, automatically making it relatively close to normal. Its kurtosis is a bit worse than the Poisson$(1)$, however.

5.3 Discrete distributions

If the $X$ is discrete, then any function $Y = g(X)$ will also be discrete, hence its pmf can be found via

$$f_Y(y) = P[Y = y] = P[g(X) = y], \quad y \in Y. \tag{5.44}$$

Of course, that final probability may or not be easy to find.

One situation in which it is easy is when $g$ is a one-to-one and onto function from $X$ to $Y$, so that there exists an inverse function,

$$g^{-1} : Y \to X; \quad g(g^{-1}(y)) = y \quad \text{and} \quad g^{-1}(g(x)) = x. \tag{5.45}$$

Then, with $f_X$ being the pmf of $X$,

$$f_Y(y) = P[g(X) = y] = P[X = g^{-1}(y)] = f_X(g^{-1}(y)). \tag{5.46}$$

For example, if $X \sim \text{Poisson}(\lambda)$, and $Y = X^2$, then $g(x) = x^2$, hence $g^{-1}(y) = \sqrt{y}$ for $y \in Y = \{0, 1, 4, 9, \ldots \}$. The pmf of $Y$ is then

$$f_Y(y) = f_X(\sqrt{y}) = e^{-\lambda} \frac{\lambda^{\sqrt{y}}}{\sqrt{y}!}, \quad y \in Y. \tag{5.47}$$

Notice that it is important to have the spaces correct. E.g., this $g$ is not one-to-one if the space is $\mathbb{R}$, and the $"\sqrt{y}"$ makes sense only if $y$ is the square of a nonnegative integer.

We consider some more examples.

5.3.1 Example: Sum of discrete uniforms

Suppose $\underline{X} = (X_1, X_2)$, where $X_1$ and $X_2$ are independent, and

$$X_1 \sim \text{Uniform} \{0, 1\} \quad \text{and} \quad X_2 \sim \text{Uniform} \{0, 1, 2\}. \tag{5.48}$$
We are after the distribution of \( Y = X_1 + X_2 \). The space of \( Y \) can be seen to be \( \mathcal{Y} = \{0, 1, 2, 3\} \). This function is not one-to-one, e.g., there are two \( x \)'s that sum to 1: \((0, 1)\) and \((1, 0)\). This is a small enough example that we can just write out all the possibilities:

\[
\begin{align*}
    f_Y(0) &= P[X_1 + X_2 = 0] = P[X = (0, 0)] = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6} \\
    f_Y(1) &= P[X_1 + X_2 = 1] = P[X = (0, 1) \text{ or } X = (1, 0)] = \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} = \frac{2}{6} \\
    f_Y(2) &= P[X_1 + X_2 = 2] = P[X = (0, 2) \text{ or } X = (1, 1)] = \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} = \frac{2}{6} \\
    f_Y(3) &= P[X_1 + X_2 = 3] = P[X = (1, 2)] = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6} 
\end{align*}
\] (5.49)

5.3.2 Example: Convolutions

Here we generalize the previous example a bit by assuming \( X_1 \) has pmf \( f_x \) and space \( \mathcal{X}_1 = \{0, 1, \ldots, a\} \), and \( X_2 \) has pmf \( f_2 \) and space \( \mathcal{X}_2 = \{0, 1, \ldots, b\} \). Both \( a \) and \( b \) are positive integers, or either could be \( +\infty \). Then \( Y \) has space

\[ \mathcal{Y} = \{0, 1, \ldots, a + b\}. \] (5.50)

To find \( f_Y(y) \), we need to sum up the probabilities of all \( (x_1, x_2) \)'s for which \( x_1 + x_2 = y \). These pairs can be written \( (x_1, y - x_1) \), and require \( x_1 \in \mathcal{X}_1 \) as well as \( y - x_1 \in \mathcal{X}_2 \). That is, for fixed \( y \in \mathcal{Y} \),

\[ 0 \leq x_1 \leq a \text{ and } 0 \leq y - x_1 \leq b \implies \max\{0, y - b\} \leq x_1 \leq \min\{a, y\}. \] (5.51)

For example, with \( a = 3 \) and \( b = 3 \), the following table shows which \( x_1 \)'s correspond to each \( y \):

<table>
<thead>
<tr>
<th>( x_1 ) ↓</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

(5.52)

Each value of \( y \) appears along a diagonal, so that

\[
\begin{align*}
    y = 0 & \Rightarrow x_1 = 0 \\
    y = 1 & \Rightarrow x_1 = 0, 1 \\
    y = 2 & \Rightarrow x_1 = 0, 1, 2 \\
    y = 3 & \Rightarrow x_1 = 0, 1, 2, 3 \\
    y = 4 & \Rightarrow x_1 = 1, 2, 3 \\
    y = 5 & \Rightarrow x_1 = 2, 3 \\
    y = 6 & \Rightarrow x_1 = 3
\end{align*}
\] (5.53)
Thus in general, for \( y \in \mathcal{Y} \),

\[
    f_Y(y) = P[X_1 + X_2 = y] = \min\{a, y\} \sum_{x_1 = \max\{0, y-b\}} P[X_1 = x_1, X_2 = y-x_1] = \min\{a, y\} \sum_{x_1 = \max\{0, y-b\}} f_1(x_1)f_2(y-x_1). \tag{5.54}
\]

This formula is often called the convolution of \( f_1 \) and \( f_2 \). In general, the convolution of two random variables is the distribution of the sum.

To illustrate, suppose \( X_1 \) has the pmf of the \( \mathcal{Y} \) in (5.49), and \( X_2 \) is \( \text{Uniform}\{0, 1, 2, 3\} \), so that \( a = b = 3 \) and

\[
    f_1(0) = f_1(3) = \frac{1}{6}, \quad f_1(1) = f_1(2) = \frac{1}{3} \quad \text{and} \quad f_2(x_2) = \frac{1}{4}, \quad x_2 = 0, 1, 2, 3. \tag{5.55}
\]

Then \( \mathcal{Y} = \{0, 1, \ldots, 6\} \), and

\[
    f_Y(0) = \sum_{x_1=0}^{0} f_1(x_1)f_2(0-x_1) = \frac{1}{24}
\]

\[
    f_Y(1) = \sum_{x_1=0}^{1} f_1(x_1)f_2(1-x_1) = \frac{1}{24} + \frac{1}{12} = \frac{3}{24}
\]

\[
    f_Y(2) = \sum_{x_1=0}^{2} f_1(x_1)f_2(1-x_1) = \frac{1}{24} + \frac{1}{12} + \frac{1}{12} = \frac{5}{24}
\]

\[
    f_Y(3) = \sum_{x_1=0}^{3} f_1(x_1)f_2(1-x_1) = \frac{1}{24} + \frac{1}{12} + \frac{1}{12} + \frac{1}{24} = \frac{6}{24}
\]

\[
    f_Y(4) = \sum_{x_1=1}^{3} f_1(x_1)f_2(1-x_1) = \frac{1}{12} + \frac{1}{12} + \frac{1}{24} = \frac{5}{24}
\]

\[
    f_Y(5) = \sum_{x_1=2}^{3} f_1(x_1)f_2(1-x_1) = \frac{1}{12} + \frac{1}{24} = \frac{3}{24}
\]

\[
    f_Y(6) = \sum_{x_1=3}^{3} f_1(x_1)f_2(1-x_1) = \frac{1}{24}. \tag{5.56}
\]

Check that the \( f_Y(y) \)'s do sum to 1.

### 5.3.3 Example: Sum of two Poissons

An example for which \( a = b = \infty \) has \( X_1 \) and \( X_2 \) independent Poisson’s, with parameters \( \lambda_1 \) and \( \lambda_2 \), respectively. Then \( Y = X_1 + X_2 \) has space \( \mathcal{Y} = \{0, 1, \ldots\} \), the same
as the spaces of the $X_i$’s. In this case, for fixed $y$, $x_1 = 0, \ldots, y$, hence

$$f_Y(y) = \sum_{x_1=0}^{y} f_1(x_1) f_2(y - x_1)$$

$$= \sum_{x_1=0}^{y} e^{-\lambda_1} \frac{\lambda_1^{x_1}}{x_1!} e^{-\lambda_2} \frac{\lambda_2^{y-x_1}}{(y-x_1)!}$$

$$= e^{-\lambda_1 - \lambda_2} \sum_{x_1=0}^{y} \frac{1}{x_1!(y-x_1)!} \lambda_1^{x_1} \lambda_2^{y-x_1}$$

$$= e^{-\lambda_1 - \lambda_2} \frac{1}{y!} \sum_{x_1=0}^{y} \frac{y!}{x_1!(y-x_1)!} \lambda_1^{x_1} \lambda_2^{y-x_1}$$

$$= e^{-\lambda_1 - \lambda_2} \frac{1}{y!} (\lambda_1 + \lambda_2)^y,$$

(5.57)

the last step using the Binomial Theorem in (3.80). But that last expression is the Poisson pmf, i.e.,

$$Y \sim \text{Poisson}(\lambda_1 + \lambda_2).$$

(5.58)

This fact can be proven also using mgf’s.

### 5.3.4 Binomials and Bernoullis

The distribution of a random variable that takes on just the values 0 and 1 is completely specified by giving the probability it is 1. Such a variable is called Bernoulli.

**Definition 3.** If $Z$ has space $\{0, 1\}$, then it is Bernoulli with parameter $p = P[Z = 1]$, written

$$Z \sim \text{Bernoulli}(p).$$

(5.59)

The pmf can then be written

$$f(z) = p^z (1 - p)^{1-z}.$$  

(5.60)

Note that $\text{Bernoulli}(p) = \text{Binomial}(1, p)$. A better alternative definition\(^1\) for the binomial is based on Bernoullis.

**Definition 4.** If $Z_1, \ldots, Z_n$ are iid Bernoulli($p$), then $X = Z_1 + \cdots + Z_n$ is binomial with parameters $n$ and $p$, written

$$X \sim \text{Binomial}(n, p).$$

(5.61)

The Bernoullis represent the individual trials, and a “1” indicates success. The moments, and mgf, of a Bernoulli are easy to find. In fact, because $Z^k = Z$ for $k = 1, 2, \ldots$,

$$E[Z] = E[Z^2] = 0 \times (1 - p) + 1 \times p = p \Rightarrow \text{Var}[Z] = p - p^2 = p(1 - p),$$

(5.62)

and

$$M_Z(t) = E[e^{tZ}] = e^0 (1 - p) + e^t p = pe^t + 1 - p.$$  

(5.63)

\(^1\)Alternative to the definition via the pmf.
Thus for $X \sim \text{Binomial}(n,p)$,

$$E[X] = nE[Z_i] = np, \text{Var}[X] = n\text{Var}[Z_i] = np(1-p),$$  \hspace{1cm} (5.64)

and

$$M_X(t) = E[e^{t(Z_1 + \cdots + Z_n)}] = E[e^{tZ_1}] \cdots E[e^{tZ_n}] = M_Z(t)^n = (pe^t + 1 - p)^n. \hspace{1cm} (5.65)$$

This mgf is the same as we found in (3.81), meaning we really are defining the same binomial.
Chapter 6

Transformations II: Continuous distributions

The mgf approach is very convenient, especially when dealing with sums of independent variables, but does have the drawback that one needs to recognize the resulting mgf (although there are ways to find the density from the mgf). Densities are typically most useful, but to find the pdf of a transformation, one cannot simply plug $g^{-1}(y)$ into the pmf of $X$, as in (5.47). The problem is that the for a pdf, $f_X(x)$ is not $P[X = x]$, which is 0. Rather, for a small area $A \subset X$ that contains $x$,

$$P[X \in A] \approx f_X(x) \times \text{Area}(A).$$  \hspace{1cm} (6.1)

Now suppose $g : \mathcal{X} \rightarrow \mathcal{Y}$ is one-to-one and onto. Then for $y \in B \subset \mathcal{Y}$,

$$P[Y \in B] \approx f_Y(y) \times \text{Area}(B).$$  \hspace{1cm} (6.2)

Now

$$P[Y \in B] = P[X \in g^{-1}(B)] \approx f_X(g^{-1}(y)) \times \text{Area}(g^{-1}(B)),$$ \hspace{1cm} (6.3)

where

$$g^{-1}(B) = \{x \in \mathcal{X} \mid g(x) \in B\},$$ \hspace{1cm} (6.4)

hence

$$f_Y(y) \approx f_X(g^{-1}(y)) \times \frac{\text{Area}(g^{-1}(B))}{\text{Area}(B)}. \hspace{1cm} (6.5)$$

Compare this equation to (5.46). For continuous distributions, we need to take care of the transformation of areas as well as the transformation of $y$ itself. The actual pdf of $Y$ is found by shrinking the $B$ in (6.5) down to the point $y$.

6.1 One dimension

When $X$ and $Y$ are random variables, the ratio of areas in (6.5) is easy to find. Because $g$ and $g^{-1}$ are one-to-one, they must be either strictly increasing or strictly decreasing.
For \( y \in Y \), take \( \epsilon \) small enough that \( B_\epsilon \equiv [y, y + \epsilon) \in Y \). Then, since the area of an interval is just its length,

\[
\frac{\text{Area}(g^{-1}(B_\epsilon))}{\text{Area}(B_\epsilon)} = \frac{|g^{-1}(y + \epsilon) - g^{-1}(y)|}{\epsilon} \\
\rightarrow \left| \frac{\partial}{\partial y} g^{-1}(y) \right| \quad \text{as} \quad \epsilon \to 0. \quad (6.6)
\]

(The absolute value is there in case \( g \) is decreasing.) That derivative is called the Jacobian of the transformation \( g^{-1} \). That is,

\[
f_Y(y) = f_X(g^{-1}(y)) \left| J_{g^{-1}}(y) \right|, \quad \text{where} \quad J_{g^{-1}}(y) = \frac{\partial}{\partial y} g^{-1}(y). \quad (6.7)
\]

This approach needs a couple of assumptions: \( y \) is in the interior of \( Y \), and the derivative of \( g^{-1}(y) \) exists.

Reprising Example 5.2.1, where \( X \sim \text{Uniform}(0, 1) \) and \( Y = -\log(X) \), we have \( Y = (0, \infty) \), and \( g^{-1}(y) = e^{-y} \). Then

\[
f_X(e^{-y}) = 1 \quad \text{and} \quad J_{g^{-1}}(y) = \frac{\partial}{\partial y} e^{-y} = -e^{-y} \implies f_Y(y) = 1 \times |e^{-y}| = e^{-y}, \quad (6.8)
\]

which is indeed the answer from (5.25).

### 6.2 General case

For the general case (for vectors), we need to figure out the ratio of the volumes, which is again given by the Jacobian, but for vectors.

**Definition 5.** Suppose \( g : \mathcal{X} \to \mathcal{Y} \) is one-to-one and onto, where both \( \mathcal{X} \) and \( \mathcal{Y} \) are open subsets of \( \mathbb{R}^p \), and all the first partial derivative of \( g^{-1}(y) \) exist. Then the Jacobian of the transformation \( g^{-1} \) is defined to be

\[
J_{g^{-1}} : \mathcal{Y} \to \mathbb{R}, \quad (6.9)
\]

\[
J_{g^{-1}}(y) = \begin{vmatrix}
\frac{\partial}{\partial y_1} g_1^{-1}(y) & \frac{\partial}{\partial y_2} g_1^{-1}(y) & \cdots & \frac{\partial}{\partial y_p} g_1^{-1}(y) \\
\frac{\partial}{\partial y_1} g_2^{-1}(y) & \frac{\partial}{\partial y_2} g_2^{-1}(y) & \cdots & \frac{\partial}{\partial y_p} g_2^{-1}(y) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial}{\partial y_1} g_p^{-1}(y) & \frac{\partial}{\partial y_2} g_p^{-1}(y) & \cdots & \frac{\partial}{\partial y_p} g_p^{-1}(y)
\end{vmatrix}, \quad (6.10)
\]

where \( g^{-1}(y) = (g_1^{-1}(y), \ldots, g_p^{-1}(y)) \), and here the “| |” represents the determinant.

The next theorem is from advanced calculus.
Theorem 2. Suppose the conditions in Definition 5, and \( J_{g^{-1}}(y) \) is continuous and non-zero for \( y \in \mathcal{Y} \). Then
\[
f_Y(y) = f_X(x) \times |J_{g^{-1}}(y)|.
\] (6.11)
Here the "\(| \cdot |\)" represents the absolute value.

If you think of \( g^{-1} \) as \( x \), you can remember the formula as
\[
f_Y(y) = f_X(x) \times \left| \frac{dx}{dy} \right|.
\] (6.12)

6.2.1 Example: Gamma – Beta

Suppose \( X_1 \) and \( X_2 \) are independent, with
\[
X_1 \sim \text{Gamma}(\alpha, \lambda) \quad \text{and} \quad X_2 \sim \text{Gamma}(\beta, \lambda),
\] (6.13)
so that they have the same scale but possibly different shapes. We are interested in
\[
Y_1 = \frac{X_1}{X_1 + X_2}.
\] (6.14)

This variable arises, e.g., in linear regression, where \( R^2 \) has that distribution under certain conditions. The function taking \((x_1, x_2)\) to \( y_1 \) is not one-to-one. To fix that up, we introduce another variable, \( Y_2 \), so that the function to the pair \((y_1, y_2)\) is one-to-one. Then to find the pdf of \( Y_1 \), we integrate out \( Y_2 \). We will take
\[
Y_2 = X_1 + X_2.
\] (6.15)

Then
\[
\mathcal{X} = (0, \infty) \times (0, \infty) \quad \text{and} \quad \mathcal{Y} = (0, 1) \times (0, \infty).
\] (6.16)

To find \( g^{-1} \), solve the equation \( y = g(x) \) for \( x \):
\[
y_1 = \frac{x_1}{x_1 + x_2}, y_2 = x_1 + x_2 \implies x_1 = y_1 y_2, x_2 = y_2 - x_1
\]
\[
\implies x_1 = y_1 y_2, x_2 = y_2(1 - y_1),
\] (6.17)

hence
\[
g^{-1}(y_1, y_2) = (y_1 y_2, y_2(1 - y_1)).
\] (6.18)

Using this inverse, we can see that indeed the function is onto the \( \mathcal{Y} \) in (6.16), since any \( y_1 \in (0, 1) \) and \( y_2 \in (0, \infty) \) will yield, via (6.18), \( x_1 \) and \( x_2 \) in \((0, \infty)\). For the Jacobian:
\[
J_{g^{-1}}(y) = \begin{vmatrix} \frac{\partial}{\partial y_1} y_1 y_2 & \frac{\partial}{\partial y_2} y_1 y_2 \\ \frac{\partial}{\partial y_1} y_2(1 - y_1) & \frac{\partial}{\partial y_2} y_2(1 - y_1) \end{vmatrix}
\]
\[
= \begin{vmatrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{vmatrix}
\]
\[
= y_2(1 - y_1) + y_1 y_2
\]
\[
= y_2.
\] (6.19)
Now because the \(X_i\)'s are independent Gamma's, their pdf is
\[
f_X(x_1, x_2) = \frac{\lambda^a}{\Gamma(a)} x_1^{a-1} e^{-\lambda x_1} \frac{\lambda^\beta}{\Gamma(\beta)} x_2^{\beta-1} e^{-\lambda x_2}.
\] (6.20)

Then the pdf of \(Y\) is
\[
f_Y(y) = f_X(g^{-1}(y)) \times |g^{-1}(y)|
= f_X(y_1 y_2, y_2(1 - y_1)) \times |y_2|
= \frac{\lambda^a}{\Gamma(a)} (y_1 y_2)^{a-1} e^{-\lambda y_1 y_2} \frac{\lambda^\beta}{\Gamma(\beta)} (y_2(1 - y_1))^{\beta-1} e^{-\lambda y_2(1 - y_1)} \times |y_2|
= \frac{\lambda^{a+\beta}}{\Gamma(a) \Gamma(\beta)} y_1^{a-1}(1 - y_1)^{\beta-1} y_2^{a+\beta-1} e^{-\lambda y_2}.
\] (6.21)

To find the pdf of \(Y_1\), we can integrate out \(y_2\):
\[
f_{Y_1}(y_1) = \int_0^\infty f_Y(y). \tag{6.22}
\]

That certainly is a fine approach. But in this case, note that the joint pdf in (6.21) can be factored into a function of just \(y_1\) and a function of just \(y_2\). That fact, coupled with the fact that the space is a rectangle, means that we automatically have that \(Y_1\) and \(Y_2\) are independent, and also have the forms of their marginal pdf's. The following lemma is more explicit.

**Lemma 8.** Suppose \(Y = (Y^{(1)}, Y^{(2)})\) has pdf \(f_Y(y) = c \ h_1(y^{(1)}) h_2(y^{(2)})\) for some constant \(c\) and functions (not necessarily pdf's) \(h_1\) and \(h_2\), and has joint space \(Y = Y_1 \times Y_2\). Then \(Y^{(1)}\) and \(Y^{(2)}\) are independent, and have marginal pdf's
\[
f_1(y^{(1)}) = c_1 \ h_1(y^{(1)}) \quad \text{and} \quad f_2(y^{(2)}) = c_2 \ h_2(y^{(2)}), \tag{6.23}
\]
respectively, where \(c = c_1 c_2\), and \(c_1\) and \(c_2\) are constants that ensure the pdf's integrate to 1.

Apply this lemma to (6.21), where
\[
c = \frac{\lambda^{a+\beta}}{\Gamma(a) \Gamma(\beta)}, \quad h_1(y_1) = y_1^{a-1}(1 - y_1)^{\beta-1} \quad \text{and} \quad h_2(y_2) = y_2^{a+\beta-1} e^{-\lambda y_2}. \tag{6.24}
\]
The lemma shows that \(Y_1\) and \(Y_2\) are independent, and the pdf of \(Y_2\) is \(c_2 h_2(y_2)\) for some \(c_2\). But notice that is the Gamma(\(\alpha + \beta, \lambda\)) pdf, hence we already know that
\[
c_2 = \frac{\lambda^{a+\beta}}{\Gamma(a + \beta)}. \tag{6.25}
\]

Then the pdf of \(Y_1\) is \(c_1 h_1(y_1)\), where it must be that \(c_1 = c/c_2\), hence
\[
f_{Y_1}(y_1) = \frac{\lambda^{a+\beta}}{\Gamma(a) \Gamma(\beta)} \frac{\Gamma(a + \beta)}{\lambda^{a+\beta}} y_1^{a-1}(1 - y_1)^{\beta-1} = \frac{\Gamma(a + \beta)}{\Gamma(a) \Gamma(\beta)} y_1^{a-1}(1 - y_1)^{\beta-1}. \tag{6.26}
\]
That is the $\text{Beta}(\alpha, \beta)$ pdf. Note that we have surreptitiously also proven that
\[
\int_0^1 y_1^{\alpha-1}(1 - y_1)^{\beta-1} dy_1 = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} = \beta(\alpha, \beta), \tag{6.27}
\]
which is the \textbf{beta function}.

To summarize, we have shown that $X_1/(X_1 + X_2)$ and $X_1 + X_2$ are independent (even though they do not really look independent), and that
\[
\frac{X_1}{X_1 + X_2} \sim \text{Beta}(\alpha, \beta) \quad \text{and} \quad X_1 + X_2 \sim \text{Gamma}(\alpha + \beta, \lambda). \tag{6.28}
\]
That last fact we already knew, from Example 5.2.2. Also, notice that the Beta variable does not depend on the scale $\lambda$.

Turning to the moments of the Beta, it is easy enough to find the $m$ directly by integrating. The mgf is not particularly helpful, as it is a confluent hypergeometric function. But since we know the first two moments of the Gamma, we can use the independence of $Y_1$ and $Y_2$ to derive them. We will take $\lambda = 1$ for simplicity, which is fine because the distribution of $Y_1$ is independent of $\lambda$. The mean and variance of a $\text{Gamma}(\alpha, 1)$ are both $\alpha$ from (3.75) and (3.76), hence
\[
E[X_1] = \alpha, \quad E[X_1^2] = \text{Var}[X_1] + E[X_1]^2 = \alpha + \alpha^2, \\
E[Y_2] = \alpha + \beta, \quad E[Y_2^2] = \alpha + \beta + (\alpha + \beta)^2. \tag{6.29}
\]
Then since $X_1 = Y_1 Y_2$ and $Y_1$ and $Y_2$ are independent,
\[
E[X_1] = E[Y_1]E[Y_2] \quad \text{and} \quad E[X_1^2] = E[Y_1^2]E[Y_2^2] \Rightarrow \alpha = E[Y_1](\alpha + \beta) \quad \text{and} \quad \alpha + \alpha^2 = E[Y_1^2](\alpha + \beta + (\alpha + \beta)^2) \\
\Rightarrow E[Y_1] = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad E[Y_1^2] = \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)}, \tag{6.30}
\]
hence
\[
\text{Var}[Y_1] = \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)} - \left( \frac{\alpha}{\alpha + \beta} \right)^2 = \frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{6.31}
\]

\subsection*{6.2.2 Example: The Dirichlet distribution}

The Dirichlet distribution is a multivariate version of the beta. We start with the independent random variables $X_1, \ldots, X_K$, where
\[
X_k \sim \text{Gamma}(\alpha_k, 1). \tag{6.32}
\]
Then the $K - 1$ vector $Y$ defined via
\[
Y_k = \frac{X_k}{X_1 + \cdots + X_K}, \quad k = 1, \ldots, K - 1, \tag{6.33}
\]
has a Dirichlet distribution, written
\[
Y \sim \text{Dirichlet}(a_1, \ldots, a_K). \tag{6.34}
\]
Chapter 6. Transformations II: Continuous distributions

There is also a $Y_K$, which may come in handy, but because $Y_1 + \cdots + Y_K = 1$, it is redundant. Also, as in the Beta, the definition is the same if the $X_k$’s are $\text{Gamma}(\alpha_k, \lambda)$.

The representation (6.33) makes it easy to find the marginals, i.e.,

$$Y_k \sim \text{Beta}(\alpha_k, 1 - \alpha_k), \quad (6.35)$$

hence the marginal means and variances. Sums of the $Y_k$’s are also beta, e.g., if $K = 4$, then

$$Y_1 + Y_3 = \frac{X_1 + X_3}{X_1 + X_2 + X_3 + X_4} \sim \text{Beta}(\alpha_1 + \alpha_3, \alpha_2 + \alpha_4). \quad (6.36)$$

The space of $Y$ is

$$Y = \{y \in \mathbb{R}^{K-1} \mid 0 < y_k < 1, k = 1, \ldots, K-1, \text{ and } y_1 + \cdots + y_{K-1} < 1\}. \quad (6.37)$$

To find the pdf of $\underline{Y}$, we need a one-to-one transformation from $\underline{X}$, so we need to append another variable to $\underline{Y}$. The easiest choice is

$$W = X_1 + \ldots + X_K, \quad (6.38)$$

so that $g(x) = (y, w)$. Then $g^{-1}$ is given by

$$x_1 = wy_1$$
$$x_2 = wy_2$$
$$\vdots$$
$$x_{K-1} = wy_{K-1}$$
$$x_K = w(1 - y_1 - \cdots - y_{K-1}). \quad (6.39)$$

Left to the reader: The Jacobian is $w^{K-1}$, and the joint pdf of $(\underline{Y}, W)$ is

$$f_{(\underline{Y}, W)}(y, w) = \left[\prod_{k=1}^{K} \frac{1}{\Gamma(\alpha_k)}\right] \left[\prod_{k=1}^{K-1} y_k^{\alpha_k-1}\right] (1 - y_1 - \cdots - y_{K-1})^{\alpha_K} w^{\alpha_1 + \cdots + \alpha_{K-1}} e^{-w}. \quad (6.40)$$

The joint space of $(\underline{Y}, W)$ can be shown to be $Y \times (0, \infty)$, which, together with the factorization of the density in (6.40), means that $Y$ and $W$ are independent. In addition, $W \sim \Gamma(\alpha_1 + \cdots + \alpha_K, 1)$, which can be seen either by looking at the $w$-part in (6.40), or by noting that it is the sum of independent gamma’s with the same scale parameter. Either way, we then have that the constant that goes with the $w$-part is $1/\Gamma(\alpha_1 + \cdots + \alpha_K)$, hence the pdf of $\underline{Y}$ must be

$$f_{\underline{Y}}(y) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \cdot y_1^{\alpha_1-1} \cdots y_{K-1}^{\alpha_{K-1}} (1 - y_1 - \cdots - y_{K-1})^{\alpha_K-1}. \quad (6.41)$$

Comparing this to the beta pdf in (6.26), we see that the Dirichlet is indeed an extension, and

$$\text{Beta}(\alpha, \beta) = \text{Dirichlet}(\alpha, \beta). \quad (6.42)$$
6.2.3 Example: Affine transformations

Suppose $X$ is $p \times 1$, and for given $p \times 1$ vector $a$ and $p \times p$ matrix $B$, let

$$ Y = g(X) = a + BX. \quad (6.43) $$

In order for this transformation to be one-to-one, we need that $B$ is invertible, which we will assume, so that

$$ X = g^{-1}(Y) = B^{-1}(Y - a). \quad (6.44) $$

For a matrix $C$ and vector $z$, with $w = Cz$, it is not hard to see that

$$ \frac{\partial w_i}{\partial z_j} = C_{ij}, \quad (6.45) $$

the $ij^{th}$ element of $C$. Thus from (6.44), since the $a$ part is a constant, the Jacobian is

$$ J_{g^{-1}}(y) = |B^{-1}|. \quad (6.46) $$

The invertibility of $B$ ensures that this Jacobian is between $0$ and $\infty$.

6.2.4 Example: Bivariate Normal

We apply this result to the normal, where $X_1$ and $X_2$ are iid $N(0, 1)$, $a$ is a $2 \times 1$ vector, and $B$ is a $2 \times 2$ invertible matrix. Then $Y$ is given as in (6.44). The mean and covariance matrix for $X$ are

$$ E[X] = 0_2 \quad \text{and} \quad \text{Cov}[X] = I_2, \quad (6.47) $$

so that from (3.96) and (3.102),

$$ \mu \equiv E[Y] = a \quad \text{and} \quad \Sigma \equiv \text{Cov}[Y] = B \text{Cov}[X] B' = BB'. \quad (6.48) $$

The space of $X$, hence of $Y$, is $\mathbb{R}^2$. To find the pdf of $Y$, we start with that of $X$:

$$ f_X(x) = \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} x_i^2} = \frac{1}{2\pi} e^{-\frac{1}{2} x'x}. \quad (6.49) $$

Then

$$ f_Y(y) = f_X(B^{-1}(y - a)) \text{abs}(|B^{-1}|) $$

$$ = \frac{1}{2\pi} e^{-\frac{1}{2} (y - a)'(BB')^{-1}(y - a)} \text{abs}(|B^{-1}|). \quad (6.50) $$

Using (6.48), we can write

$$ (B^{-1}(y - a))'B^{-1}(y - a) = (y - \mu)'(BB')^{-1}(y - \mu) = (y - \mu)'\Sigma^{-1}(y - \mu), \quad (6.51) $$

and using properties of determinants,

$$ \text{abs}(|B^{-1}|) = \sqrt{|B^{-1}|} = \sqrt{|(B|B'|)^{-1}|} = \sqrt{|\Sigma|}, \quad (6.52) $$
hence the pdf of $Y$ can be given as a function of the mean and covariance matrix:

$$f_Y(y) = \frac{1}{2\pi} \frac{1}{\sqrt{|\Sigma|}} e^{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)}. \quad (6.53)$$

This $Y$ is bivariate normal, with mean $\mu$ and covariance matrix $\Sigma$, written very much in the same way as the regular Normal,

$$Y \sim N(\mu, \Sigma). \quad (6.54)$$

In particular, the $X$ here is

$$X \sim N(0, I). \quad (6.55)$$

The mgf of a bivariate normal is not hard to find given that of the $X$. Because $X_1$ and $X_2$ are iid $N(0, 1)$, their mgf is

$$M_X(t) = e^{\frac{1}{2}t_1^2}e^{\frac{1}{2}t_2^2} = e^{\frac{1}{2}\|t\|^2}. \quad (6.56)$$

Then with $Y = a + BX$,

$$M_Y(t) = E[e^{t'Y}] = E[e^{t'(a + BX)}] = e^{t'a}E[e^{t'BX}] = e^{t'a}M_X(B't) = e^{t'a}e^{\frac{1}{2}\|B't\|^2} = e^{t'a+\frac{1}{2}t'St} \text{ by (6.48),} \quad (6.57)$$

because $\|B't\|^2 = (B't)'B't = t'BB't = t'St$. Compare this mgf to that of the regular normal, in (3.65).

Later, we will deal with the general $p \times 1$ multivariate normal, proceeding exactly the same way as above, except the matrices and vectors have more elements.

### 6.2.5 Example: Orthogonal transformations

A $p \times p$ orthogonal matrix is a matrix $\Gamma$ such that

$$\Gamma'\Gamma = \Gamma\Gamma' = I_p. \quad (6.58)$$

An orthogonal matrix has orthonormal columns (and orthonormal rows), that is, with $\Gamma = (\gamma_1, \ldots, \gamma_p)$,

$$\|\gamma_i\| = 1 \quad \text{and} \quad \gamma_i'\gamma_j = 0 \quad \text{if} \quad i \neq j. \quad (6.59)$$

An orthogonal transformation of $X$ is then

$$Y = \Gamma X \quad (6.60)$$

for some orthogonal matrix $\Gamma$. This transformation rotates $Y$ about zero, that is, the length of $Y$ is the same as that of $X$, but its orientation is different, because

$$\|\Gamma z\|^2 = z'\Gamma'\Gamma z = z'z = \|z\|^2. \quad (6.61)$$

The Jacobian is $\pm 1$:

$$|\Gamma'| = |I_p| = 1 \implies |\Gamma|^2 = 1. \quad (6.62)$$

Thus in the bivariate Normal example above if $X \sim N(0, I)$, then so is $Y = \Gamma X$. 
6.2. General case

6.2.6 Example: Polar coordinates

Suppose \( \mathbf{X} = (X_1, X_2) \) has pdf \( f_X(x) \) for \( x \in \mathcal{X} \), and consider the transformation to polar coordinates

\[
g(x_1, x_2) = (r, \theta) = (\|\mathbf{x}\|, \text{Angle}(x_1, x_2)), \quad \text{where } \|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2}. \tag{6.63}
\]

The \( \text{Angle}(x_1, x_2) \) is taken to be in \([0, 2\pi)\), and is basically the \( \arctan(x_2/x_1) \), except that that is not uniquely defined, e.g., if \( x_1 = x_2 = 1 \), \( \arctan(1/1) = \pi/4 \) or \( 5\pi/4 \). What we really mean is the unique value \( \theta \) in \([0, 2\pi)\) for which

\[
x_1 = r \cos(\theta) \quad \text{and} \quad x_2 = r \sin(\theta). \tag{6.64}
\]

Another glitch is that \( \theta \) is not uniquely defined when \( (x_1, x_2) = (0, 0) \), since then \( r = 0 \) and any \( \theta \) would work. So we should assume that \( \mathcal{X} \) does not contain \( (0, 0) \).

Just to be technical, we also need the space of \( (r, \theta) \) to be open (from Theorem 2), so we should just go ahead and assume \( x_2 \neq 0 \) for \( (x_1, x_2) \in \mathcal{X} \). This requirement would not hurt anything because we are dealing with continuous random variables, but for simplicity we will ignore these niceties.

We can find the Jacobian without knowing the pdf or space of \( \mathbf{X} \), and the \( g^{-1} \) is already given in (6.64), hence

\[
J_{g^{-1}}(r, \theta) = \begin{vmatrix}
\frac{\partial}{\partial r} r \cos(\theta) & \frac{\partial}{\partial \theta} r \cos(\theta) \\
\frac{\partial}{\partial r} r \sin(\theta) & \frac{\partial}{\partial \theta} r \sin(\theta)
\end{vmatrix} = \begin{vmatrix}
\cos(\theta) & -r \sin(\theta) \\
\sin(\theta) & r \cos(\theta)
\end{vmatrix} = -r \sin(\theta)^2 - r \cos(\theta)^2 = -r. \tag{6.65}
\]

You may recall this result from calculus, i.e.,

\[
dx_1 dx_2 = r dr d\theta. \tag{6.66}
\]

The pdf of \( (r, \theta) \) in any case is

\[
f_{\mathbf{Y}}(r, \theta) = f_{\mathbf{X}}(r \cos(\theta), r \sin(\theta)) r. \tag{6.67}
\]

We next look at some examples.

**Spherical symmetric pdf’s**

A random variable \( \mathbf{X} \) is spherically symmetric if for any orthogonal \( \Gamma \),

\[
\mathbf{X} =^D \Gamma \mathbf{X}, \tag{6.68}
\]

meaning \( \mathbf{X} \) and \( \Gamma \mathbf{X} \) have the same distribution. For example, \( \mathbf{X} \sim N(0, I_2) \) is spherically symmetric.
Suppose the $2 \times 1$ vector $\mathbf{X}$ is spherically symmetric and has a pdf $f_X(x)$. Let $\mathcal{R}$ be the space of $R \equiv \|\mathbf{X}\|$. Then the space of $\mathbf{X}$ must include all circles with radii in $\mathcal{R}$, that is,

$$\mathcal{X} = \bigcup_{r \in \mathcal{R}} \{ \mathbf{x} \mid x_1 + x_2 = r \}. \quad (6.69)$$

In addition, the pdf of $\mathbf{X}$ depends on $\mathbf{x}$ through only $r$, that is,

$$f_X(x) = h(||\mathbf{x}||) \quad (6.70)$$

for some function $h : [0, \infty) \to [0, \infty)$. This $h$ is not itself a pdf. (We can see that if $Y = \Gamma Y$, then $f_Y(y) = f_X(\Gamma' y) = h(||\Gamma'y||) = h(||y||) = f_X(y)$, (6.71), i.e., $Y$ has the same pdf as $X$.)

Now with $y = (r, \text{Angle}(x_1, x_2))$, (6.69) shows that the $\mathcal{Y} = \mathcal{R} \times [0, 2\pi)$, (6.72) and (6.67) shows that

$$f_Y(r, \theta) = f_X(r \cos(\theta), r \sin(\theta))r = h(r)r, \quad (6.73)$$

because $(r \cos(\theta))^2 + (r \sin(\theta))^2 = r^2$.

Now we can appeal to Lemma 8 about independence. This pdf can be written as a function of just $r$, i.e., $h(r)r$, and a function of just $\theta$, i.e., the function “1”, and $c = 1$. The space is a rectangle. Thus we know that $\theta$ has pdf $c_2$ for $\theta \in [0, 2\pi)$. That is,

$$\theta \sim \text{Uniform}[0, 2\pi), \quad (6.74)$$

and $c_2 = 1/(2\pi)$. Then $c_1 = c/c_2 = 2\pi$, i.e., the marginal pdf of $R$ is

$$f_R(r) = 2\pi h(r)r. \quad (6.75)$$

Applying this approach to $\mathbf{X} \sim N(\mathbf{0}_2, \mathbf{I}_2)$, we know from (6.49) that

$$h(r) = \frac{1}{2\pi} e^{-\frac{1}{2} r^2}, \quad (6.76)$$

hence $R$ has pdf

$$f_R(r) = r e^{-\frac{1}{2} r^2}. \quad (6.77)$$

The **Box-Mueller transformation** is an approach to generating two random Normals from two random uniforms that reverses the above procedure. That is, suppose $U_1$ and $U_2$ are independent $\text{Uniform}(0, 1)$. Then we can generate $\theta$ by setting

$$\theta = 2\pi U_1, \quad (6.78)$$

and $R$ by setting

$$R = F_R^{-1}(U_2), \quad (6.79)$$

where $F_R$ is the distribution function for the pdf in (6.77):

$$F_R(r) = \int_0^r w e^{-\frac{1}{2} u^2} du = -e^{\frac{1}{2} u^2} \bigg|_{u=0}^{u=r} = 1 - e^{-\frac{1}{2} r^2}. \quad (6.80)$$
Inverting $u_2 = F_R(r)$ yields
\[ r = F^{-1}(u_2) = \sqrt{-2\log(1 - u_2)}. \] (6.81)
Thus, as in (6.64), we set
\[ X_1 = \sqrt{-2\log(1 - U_2)} \cos(2\pi U_1) \quad \text{and} \quad X_2 = \sqrt{-2\log(1 - U_2)} \sin(2\pi U_1), \] (6.82)
which are then independent $N(0, 1)$’s. (Usually one sees $U_2$ in place of the $1 - U_2$ in the logs, but either way is fine because both are $\text{Uniform}(0, 1)$.)

### 6.3 Order statistics

The order statistics for a sample \( \{x_1, \ldots, x_n\} \) are the observations placed in order from smallest to largest. They are usually designated with indices \((i)\), so that the order statistics are
\[ x_{(1)}, x_{(2)}, \ldots, x_{(n)}, \] (6.83)
where
\[ x_{(1)} = \text{smallest of} \{x_1, \ldots, x_n\} = \min\{x_1, \ldots, x_n\} \]
\[ x_{(2)} = \text{second smallest of}\{x_1, \ldots, x_n\} \]
\[ \vdots \]
\[ x_{(n)} = \text{largest of} \{x_1, \ldots, x_n\} = \max\{x_1, \ldots, x_n\}. \] (6.84)
For example, if the sample is \( \{3.4, 2.5, 1.7, 5.2\} \) then the order statistics are 1.7, 2.5, 3.4, 5.2. If two observations have the same value, then that value appears twice, i.e., the order statistics for \( \{3.4, 1.7, 1.7, 5.2\} \) are 1.7, 1.7, 3.4, 5.2, i.e., $x_{(1)} = x_{(2)}$.

These statistics are useful as descriptive statistics, and in nonparametric inference. For example, estimates of the median of a distribution are often based on order statistics, such as the median, or the trimean, which is a linear combination of the two quartiles and the median.

We will deal with $X_1, \ldots, X_n$ being iid, each with distribution function $F$, pdf $f$, and space $\mathcal{X}_0$, so that the space of $X = (X_1, \ldots, X_n)$ is $\mathcal{X} = \mathcal{X}_0 \times \cdots \times \mathcal{X}_0$. Then let
\[ Y = (X_{(1)}, \ldots, X_{(n)}). \] (6.85)
We will assume that the $X_i$’s are distinct, that is, no two have the same value. This assumption is fine in the continuous case, because the probability that two observations are equal is zero. In the discrete case, there may indeed be ties, and the analysis becomes more difficult. Thus the space of $Y$ is
\[ \mathcal{Y} = \{(y_1, \ldots, y_n) \mid y_i \in \mathcal{X}_0, y_1 < y_2 < \cdots < y_n\}. \] (6.86)
To find the pdf, start with $y \in \mathcal{Y}$, and let $\delta > 0$ be small enough so that the intervals $(y_1, y_1 + \delta), (y_2, y_2 + \delta), \ldots, (y_n, y_n + \delta)$ are disjoint. (So take $\delta$ less than the all the gaps $y_{i+1} - y_i$.) Then
\[ P[y_1 < Y_1 < y_1 + \delta, \ldots, y_n < Y_n < y_n + \delta] = P[y_1 < X_{(1)} < y_1 + \delta, \ldots, y_n < X_{(n)} < y_n + \delta]. \] (6.87)
Now the event in the latter probability occurs when any permutation of the \( x_i \)'s has one component in the first interval, one in the second, etc. E.g., if \( n = 3 \),

\[
P[y_1 < X(1) < y_1 + \delta, y_2 < X(2) < y_2 + \delta, y_3 < X(3) < y_3 + \delta] = P[y_1 < X_1 < y_1 + \delta, y_2 < X_2 < y_2 + \delta, y_3 < X_3 < y_3 + \delta] + P[y_1 < X_1 < y_1 + \delta, y_2 < X_3 < y_2 + \delta, y_3 < X_2 < y_3 + \delta] + P[y_1 < X_2 < y_1 + \delta, y_2 < X_1 < y_2 + \delta, y_3 < X_3 < y_3 + \delta] + P[y_1 < X_2 < y_1 + \delta, y_2 < X_3 < y_2 + \delta, y_3 < X_1 < y_3 + \delta] + P[y_1 < X_3 < y_1 + \delta, y_2 < X_1 < y_2 + \delta, y_3 < X_3 < y_3 + \delta] + P[y_1 < X_3 < y_1 + \delta, y_2 < X_2 < y_2 + \delta, y_3 < X_1 < y_3 + \delta] = 6 P[y_1 < X_1 < y_1 + \delta, y_2 < X_2 < y_2 + \delta, y_3 < X_3 < y_3 + \delta]. \quad (6.88)
\]

The last equation follows because the \( x_i \)'s are iid, hence the six individual probabilities are equal. In general, the number of permutations is \( n! \). Thus, we can write

\[
P[y_1 < Y_1 < y_1 + \delta, \ldots, y_n < Y_n < y_n + \delta] = n!P[y_1 < X_1 < y_1 + \delta, \ldots, y_n < X_n < y_n + \delta] = n! \prod_{i=1}^{n} [F(y_i + \delta) - F(y_i)]. \quad (6.89)
\]

Dividing by \( \delta^n \) then letting \( \delta \to 0 \) yields the joint density, which is

\[
f_{\underline{Y}}(\underline{y}) = n! \prod_{i=1}^{n} f(y_i), \quad \underline{y} \in \mathcal{Y}. \quad (6.90)
\]

Marginal distributions of individual order statistics, or sets of them, can be obtained by integrating out the ones that are not desired. The process can be a bit tricky, and one must be careful with the spaces. Instead, we will present a representation that leads to the marginals as well as other quantities.

We start with the \( U_1, \ldots, U_n \)'s being iid \( \text{Uniform}(0,1) \), so that the pdf (6.90) of the order statistics \( (U(1), \ldots, U(n)) \) is simply \( n! \). Consider the first order statistic plus the gaps between consecutive order statistics:

\[
G_1 = U(1), G_2 = U(2) - U(1), \ldots, G_n = U(n) - U(n-1). \quad (6.91)
\]

These \( G_i \)'s are all positive, and they sum to \( U(n) \), which has range \((0,1)\). Thus the space of \( \underline{G} = (G_1, \ldots, G_n) \) is

\[
\mathcal{G} = \{ \underline{g} \in \mathbb{R}^n \mid 0 < g_i < 1, i = 1, \ldots, n \ \& \ g_1 + \cdots + g_n < 1 \}. \quad (6.92)
\]

The inverse function to (6.91) is

\[
u(1) = g_1, u(2) = g_1 + g_2, \ldots, u(n) = g_1 + \cdots + g_n, \quad (6.93)
\]

so that the Jacobian is

\[
\begin{vmatrix}
1 & 0 & 0 & \cdots & 0 \\
1 & 1 & 0 & \cdots & 0 \\
1 & 1 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 1 & \cdots & 1
\end{vmatrix} = 1. \quad (6.94)
\]
Thus the pdf of $G$ is also $n!$:

$$f_G(g) = n!, \quad g \in G. \quad (6.95)$$

This pdf is quite simple on its own, but note that it is a special case of the Dirichlet in (6.41) with $K = n + 1$ and all $a_k = 1$. Thus any order statistic is a beta, because it is the sum of the first few gaps, i.e.,

$$U(k) = g_1 + \cdots + g_k \sim \text{Beta}(k, n - k + 1) \quad (6.96)$$

analogous to (6.36). In particular, if $n$ is odd, then the median of the observations is $U(n+1)/2$, hence

$$\text{Median}\{U_1, \ldots, U_n\} \sim \text{Beta}\left(\frac{n+1}{2}, \frac{n+1}{2}\right), \quad (6.97)$$

and by (6.30) and (6.31),

$$E[\text{Median}\{U_1, \ldots, U_n\}] = \frac{1}{2} \quad (6.98)$$

and

$$\text{Var}[\text{Median}\{U_1, \ldots, U_n\}] = \frac{(n+1)/2(n+1)/2}{(n+1)^2(n+2)} = \frac{1}{4(n+2)}. \quad (6.99)$$

To obtain the pdf of the order statistics of non-uniforms, we can use the probability transform approach as in Example 5.1.3. That is, suppose the $X_i$’s are iid with (strictly increasing) distribution function $F$ and pdf $f$. We then have that $X$ has the same distribution as $(F^{-1}(U_1), \ldots, F^{-1}(U_n))$, where the $U_i$’s are iid $\text{Uniform}(0,1)$. Because $F$, hence $F^{-1}$, is increasing, the order statistics for the $X_i$’s match those of the $U_i$’s, that is,

$$(X(1), \ldots, X(n)) \overset{D}{=} (F^{-1}(U(1)), \ldots, F^{-1}(U(n))). \quad (6.100)$$

Thus for any particular $k$, $X(k) \overset{D}{=} F^{-1}(U(k))$. We know that $U(k) \sim \text{Beta}(k, n - k + 1)$, hence can find the distribution of $X(k)$ using the transformation with $h(u) = F^{-1}$. Thus $h^{-1} = F$, i.e.,

$$U(k) = h^{-1}(X(k)) = F(X(k)) \Rightarrow J_{h^{-1}}(x) = |F'(x)| = f(x). \quad (6.101)$$

The pdf of $X(k)$ is then

$$f_{X(k)}(x) = f_{U(k)}(F(x))f(x) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}F(x)^{k-1}(1-F(x))^{n-k}f(x). \quad (6.102)$$

### 6.3.1 Approximating the mean and variance

For most $F$ and $k$, the pdf in (6.102) is not particularly easy to deal with analytically. If the variance of the $U(k)$ is small, one could use the “$\Delta$-method” to approximate the mean and variance of $X(k)$. This method a applied to a function $h(U)$ where $h$ is continuously differentiable ant $U = \mu_U$, its mean. We start with the one-term Taylor Series expansion about $U = \mu_U$:

$$h(U) = h(\mu_U) + (U - \mu_U)h'(\mu^*), \quad \mu^* \text{ is between } U \text{ and } \mu_U. \quad (6.103)$$
Chapter 6. Transformations II: Continuous distributions

Now if the variance of $U$ is small, then it is likely that $U$ and $\mu_U$ are fairly close, so that $\mu^*$ is close to $\mu_U$, and if $h'$ is not very variable, $h'(\mu^*) \approx h'(\mu_U)$, so that

$$h(U) \approx h(\mu_U) + (U - \mu_U)h'(\mu_U).$$

(6.104)

Because $h(\mu_U)$ and $h'(\mu_U)$ are constants, we have that

$$E[h(U)] \approx E[h(\mu_U) + (U - \mu_U)h'(\mu_U)] = h(\mu_U),$$

(6.105)

and

$$\text{Var}[h(U)] \approx \text{Var}[h(\mu_U) + (U - \mu_U)h'(\mu_U)] = (h'(\mu_U))^2 \text{Var}[U].$$

(6.106)

Later, in Section 16.5, we’ll give a more formal justification for this procedure.

To apply the $\Delta$-method to an order statistic, take $X_{(k)} = F^{-1}(U_{(k)})$, i.e., $h(u) = F^{-1}(u)$. Then one can show that

$$h'(u) = \frac{1}{f(F^{-1}(u))}.$$  

(6.107)

From (6.96), $\mu_U = k/(n + 1)$, so that

$$E[X_{(k)}] \approx F^{-1}\left(\frac{k}{n + 1}\right),$$  

(6.108)

and

$$\text{Var}[X_{(k)}] \approx \frac{1}{f(F^{-1}(\frac{k}{n+1}))^2} \left[\frac{k(n-k+1)}{(n+1)^2(n+2)}\right].$$

(6.109)

With $n$ odd, the median then has

$$E[X_{(k)}] \approx \eta \quad \text{and} \quad \text{Var}[X_{(k)}] \approx \frac{1}{f(\eta)^2} \frac{1}{4(n+2)},$$

(6.110)

where $\eta = F^{-1}\left(\frac{1}{2}\right)$ is the population median of $X$.

When the distribution of $X$ is symmetric, and the mean exists, then the mean and median are the same. Thus both the sample mean and the sample median are estimates of the mean. In terms of estimation, the better estimate would have the smaller variance. Here is a table comparing the variance of the sample mean ($\sigma^2/n$) to the approximate variance of the sample median for some symmetric distributions:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$f(\mu)$</th>
<th>$\text{Var}[\bar{X}]$</th>
<th>$\approx \text{Var}[\text{Median}]$</th>
<th>$\approx \text{Var}[\bar{X}] / \text{Var}[\text{Median}]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal($\mu, 1$)</td>
<td>$1/\sqrt{2\pi}$</td>
<td>$1/n$</td>
<td>$\pi/(2(n+2))$</td>
<td>$\frac{\pi}{n} \approx 0.6366$</td>
</tr>
<tr>
<td>Cauchy</td>
<td>$1/\pi$</td>
<td>$\infty$</td>
<td>$\pi^2/(4(n+2))$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>Double Exponential</td>
<td>$1/2$</td>
<td>$2/n$</td>
<td>$1/(n+2)$</td>
<td>$2$</td>
</tr>
<tr>
<td>Uniform($\mu - 1, \mu + 1$)</td>
<td>$1/2$</td>
<td>$1/(3n)$</td>
<td>$1/(n+2)$</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>Logistic</td>
<td>$1/4$</td>
<td>$\pi^2/(3n)$</td>
<td>$4/(n+2)$</td>
<td>$\frac{\pi^2}{12} \approx 0.8225$</td>
</tr>
</tbody>
</table>

The last column takes $n \to \infty$. So the mean is better for the Normal and Uniform, but the median is better for the others, especially the Cauchy. Note that these approximations are really for large $n$. Remember that for $n = 1$, the mean and median are the same.
Chapter 7

Conditional distributions

7.1 Introduction

A two-stage process is described in Example 2.4.3, where one first randomly chooses a coin from a population of coins, then flips it independently \( n = 10 \) times. There are two random variables in this experiment: \( X \), the probability of heads for the chosen coin, and \( Y \), the total number of heads among the \( n \) flips. It is given that \( X \) is equally likely to be any number between 0 and 1, i.e.,

\[
X \sim \text{Uniform}(0,1). \tag{7.1}
\]

Also, once the coin is chosen, \( Y \) is Binomial. If the chosen coin has \( X = x \), then we say that the conditional distribution of \( Y \) given \( X = x \) is \( \text{Binomial}(n,x) \), written

\[
Y \mid X = x \sim \text{Binomial}(n,x). \tag{7.2}
\]

Together the equations (7.1) and (7.2) describe the distribution of \( (X,Y) \).

A couple of other distributions may be of interest. First, what is the marginal, sometimes referred to in this context as unconditional, distribution of \( Y \)? It is not Binomial. It is the distribution arising from the entire two-stage procedure, not that arising given a particular coin. The space is \( \mathcal{Y} = \{0,1,\ldots,n\} \), as for the binomial, but the pmf is different:

\[
f_Y(y) = P[Y = y] \neq P[Y = y \mid X = x]. \tag{7.3}
\]

(That last expression is pronounced the probability that \( Y = y \) given \( X = x \).)

Also, one might wish to interchange the roles of \( X \) and \( Y \), and ask for the conditional distribution of \( X \) given \( Y = y \) for some \( y \). This distribution is of particular interest in Bayesian inference, as follows. One chooses a coin as before, and then wishes to know its \( X \). It is flipped ten times, and the number of heads observed, \( y \), is used to guess what the \( X \) is. More precisely, one then finds the conditional distribution of \( X \) given \( Y = y \):

\[
X \mid Y = y \sim ??? \tag{7.4}
\]

In Bayesian parlance, the marginal distribution of \( X \) in (7.1) is the prior distribution of \( X \), because it is your best guess before seeing the data, and the conditional distribution in (7.4) is the posterior distribution, determined after you have seen the data.
These ideas extend to random vectors \((X, Y)\). There are five distributions we consider, three of which we have seen before:

- **Joint.** The joint distribution of \((X, Y)\) is the distribution of \(X\) and \(Y\) taken together.
- **Marginal.** The two marginal distributions: that of \(X\) alone and that of \(Y\) alone.
- **Conditional.** The two conditional distributions: that of \(Y\) given \(X = x\), and that of \(X\) given \(Y = y\).

The next section shows how to find the joint distribution from a conditional and marginal. Further sections look at finding the marginals and reverse conditional, the latter using Bayes Theorem. We end with independence, \(Y\) begin independent of \(X\) if the conditional distribution of \(Y\) given \(X = x\) does not depend on \(x\).

### 7.2 Examples of conditional distributions

When considering the conditional distribution of \(Y\) given \(X = x\), it may or may not be that the randomness of \(X\) is of interest, depending on the situation. In addition, there is no need for \(Y\) and \(X\) to be of the same type, e.g., in the coin example, \(X\) is continuous and \(Y\) is discrete. Next we look at some additional examples.

#### 7.2.1 Simple linear model

The relationship of one variable to another is central to many statistical investigations. The simplest is a linear relationship,

\[
Y = \alpha + \beta X + E, \quad (7.5)
\]

Here, \(\alpha\) and \(\beta\) are fixed, \(Y\) is the “dependent” variable, and \(X\) is the “explanatory” or “independent” variable. The \(E\) is error, needed because one does not expect the variables to be exactly linearly related. Examples include \(X = \text{Height}\) and \(Y = \text{Weight}\), or \(X = \text{Dosage}\) of a drug and \(Y\) some measure of health (cholesterol level, e.g.). The \(X\) could be a continuous variable, or an indicator function, e.g., be 0 or 1 according to the sex of the subject.

The Normal linear model specifies that

\[
Y \mid X = x \sim N(\alpha + \beta x, \sigma^2_E). \quad (7.6)
\]

In particular,

\[
E[Y \mid X = x] = \alpha + \beta x \quad \text{and} \quad Var[Y \mid X = x] = \sigma^2_E. \quad (7.7)
\]

(Other models take (7.7) but do not assume normality, or allow \(Var[Y \mid X = x]\) to depend on \(x\).) It may be that \(X\) is fixed by the experimenter, for example, the dosage \(x\) might be preset; or it may be that the \(X\) is truly random, e.g., the height of a randomly chosen person would be random. Often, this randomness of \(X\) is ignored, and analysis proceeds conditional on \(X = x\). Other times, the randomness of \(X\) is also incorporated into the analysis, e.g., one might have the marginal

\[
X \sim N(\mu_X, \sigma^2_X). \quad (7.8)
\]
7.2. Examples of conditional distributions

7.2.2 Finite mixture model

The population may consist of a finite number of distinct subpopulations, e.g., in assessing consumer ratings of cookies, there may be a subpopulation of people who like sweetness, and one with those who do not. With \( K \) subpopulations, \( X \) takes on the values \( \{1, \ldots, K\} \). These values are indices, not necessarily meaning to convey any ordering. For a Normal mixture, the model is

\[
Y \mid X = k \sim N(\mu_k, \sigma_k^2).
\] (7.9)

Generally there are no restrictions on the \( \mu_k \)'s, but the \( \sigma_k^2 \)'s may be assumed equal.

Also, \( K \) may or may not be known. The marginal distribution for \( X \) is usually unrestricted, i.e.,

\[
f_X(k) = p_k, \quad k = 1, \ldots, K,
\] (7.10)

the only restrictions being the necessary ones, i.e., the \( p_k \)'s are positive and sum to 1.

7.2.3 Hierarchical models

Many experiments involve first randomly choosing a number of subjects from a population, the measuring a number of random variables on the chosen subjects. For example, one might randomly choose \( n \) third-grade classes from a city, then within each class administer a test to randomly choose \( m \) students. Let \( X_i \) be the overall ability of class \( i \), and \( Y_i \) the average performance on the test of the students chosen from class \( i \). Then a possible hierarchical model is

\[
X_1, \ldots, X_n \text{ are iid } \sim N(\mu, \sigma^2), \quad \text{and}
\]

\[
Y_1, \ldots, Y_n \mid X_1 = x_1, \ldots, X_n = x_n \text{ are independent } \sim \left( N(x_1, \tau^2), \ldots, N(x_n, \tau^2) \right)
\] (7.11)

Here, \( \mu \) and \( \sigma^2 \) are the mean and variance for the entire population of classes, while \( x_i \) is the mean for class \( i \). Interest may center on the overall mean, so the city can obtain funding for the state, as well as for the individual classes chosen, so these classes can get special treats from the local school board.

7.2.4 Bayesian models

A statistical model typically depends on an unknown parameter vector \( \theta \), and the objective is to estimate the parameters, or some function of them, or test hypotheses about them. The Bayesian approach treats the data \( \mathbf{X} \) and the parameter \( \theta \) as both being random, hence having a joint distribution. (See The frequentist approach considers the parameters to be fixed but unknown. Both approaches use a model for \( X_i \) which is a set of distributions indexed by the parameter, e.g., the \( X_i \)'s are iid \( N(\mu, \sigma^2) \), where \( \theta = (\mu, \sigma^2) \). The Bayesian approach considers that model conditional on \( \theta = \theta \), and would write

\[
X_1, \ldots, X_n \mid \theta = \theta \left(= (\mu, \sigma^2)\right) \sim \text{iid } N(\mu, \sigma^2).
\] (7.12)

Here, the bold \( \theta \) is the random vector, and the regular \( \theta \) is the particular value. Then to fully specify the model, the distribution of \( \theta \) must be given,

\[
\theta \sim \pi,
\] (7.13)

for some prior distribution \( \pi \). Once the data is \( \mathbf{x} \) obtained, inference is based on the posterior distribution of \( \theta \mid \mathbf{X} = \mathbf{x} \).
7.3 Conditional + Marginal → Joint

We start with the conditional distribution of $Y$ given $X = x$, and the marginal distribution of $X$, and find the joint distribution of $(X, Y)$. Let $\mathcal{X}$ be the (marginal) space of $X$, and for each $x \in \mathcal{X}$, let $\mathcal{Y}_x$ be the conditional space of $Y$ given $X = x$, that is, the space for the distribution of $Y \mid X = x$. Then the (joint) space of $(X, Y)$ is

$$\mathcal{W} = \{(x, y) \mid x \in \mathcal{X} \& y \in \mathcal{Y}_x\}. \quad (7.14)$$

(In the coin example, and $\mathcal{Y}_x = \{0, 1, \ldots, n\}$, so that in this case the conditional space of $Y$ does not depend on $x$.)

Now for a function $g(x, y)$, the conditional expectation of $g(X, Y)$ given $X = x$ is denoted

$$e_g(x) = E[g(x, Y) \mid X = x], \quad (7.15)$$

and is defined to be the expected value of the function $g(x, Y)$ where $x$ is fixed and $Y$ has the conditional distribution $Y \mid X = x$. If this conditional distribution has a pdf, say $f_{Y \mid X}(y \mid x)$, then

$$e_g(x) = \int_{\mathcal{Y}_x} g(x, y) f_{Y \mid X}(y \mid x) dy. \quad (7.16)$$

If $f_{X \mid Y}$ is a pmf, then we have the summation instead of the integral. It is important to realize that this conditional expectation is a function of $x$. In the coin example (7.2), with $g(x, y) = y$,

$$e_g(x) = E[Y \mid \theta = x] = E[\text{Binomial}(n, x)] = nx. \quad (7.17)$$

The key to describing the joint distribution is to define the unconditional expected value of $g$.

**Definition 6.** Given a function $g : \mathcal{W} \to \mathbb{R}$,

$$E[g(X, Y)] = E[e_g(X)], \quad (7.18)$$

where $e_g(x) = E[g(x, Y) \mid X = x]$, if all the expected values exist.

So, to continue with the coin example, with $g(x, y) = y$,

$$E[Y] = E[e_g(X)] = E[n\theta] = n E[\text{Uniform}(0, 1)] = \frac{n}{2}. \quad (7.19)$$

Notice that this unconditional expected value does not depend on $x$, while the conditional expected value does, which is as it should be.

This definition yields the joint distribution $P$ on $\mathcal{W}$ by looking at indicator functions $g$, as in Section 3.1.1. That is, take $A \subset \mathcal{W}$, so that with $P$ being the joint probability distribution for $(X, Y)$,

$$P[A] = E[I_A(X, Y)]. \quad (7.20)$$

Then the definition says that

$$P[A] = E[e_{I_A}(X)] \quad \text{where} \quad e_{I_A}(x) = E[I_A(x, Y) \mid X = x]. \quad (7.21)$$

We should check that this $P$ is in fact a legitimate probability distribution, and in turn yields the correct expected values. The latter result is proven in measure theory. That
it is a probability measure (see Section 2.1) is not hard to show using that $I_W \equiv 1$, and if $A_i$’s are disjoint,

$$I_{\cup A_i}(x, y) = \sum I_{A_i}(x, y). \quad (7.22)$$

The $e_{I_A}(x)$ is the conditional probability of $A$ given $X = x$ written

$$P[A \mid X = x] = P[(X, Y) \in A \mid X = x] = E[I_A(x, y) \mid X = x]. \quad (7.23)$$

### 7.3.1 Joint densities

More useful is the pdf or pmf of the joint distribution, if it exists. Suppose that $X$ is discrete, and the conditional distribution of $Y$ given $X = x$ is discrete for each $x$. The pmf’s are

$$f_X(x) = P[X = x] \quad \text{and} \quad f_{Y \mid X}(y \mid x) = P[Y = y \mid X = x]. \quad (7.24)$$

Let $f$ be the joint pdf of $(X, Y)$, so that for each $(x, y) \in W$,

$$f(x, y) = P[(X, Y) = (x, y)] = E[I_{\{(x, y)\}}(X, Y)]. \quad (7.25)$$

By (7.21),

$$E[I_{\{(x, y)\}}(X, Y)] = E[e_{I_{\{(x, y)\}}}(x)] = e_{I_{\{(x, y)\}}}(x)f_X(x), \quad (7.26)$$

where

$$e_{I_{\{(x, y)\}}}(x) = E[I_{\{(x, y)\}}(x, y) \mid X = x] = 1 \times P[Y = y \mid X = x] = f_{Y \mid X}(y \mid x). \quad (7.27)$$

Thus,

$$f(x, y) = f_{Y \mid X}(y \mid x)f_X(x). \quad (7.28)$$

In fact, this equation should not be especially surprising. It is a special case of the general definition of conditional probability for sets $A$ and $B$:

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]}, \quad \text{if} \quad P[B] \neq 0. \quad (7.29)$$

If $X$ has pdf $f_X(x)$ and $\{Y \mid X = x\}$ has pdf $f_{Y \mid X}(y \mid x)$ for each $x$, then the joint distribution has a pdf $f(x, y)$, again given by the formula in (7.28). To verify this formula, for $A \subset W$,

$$P[A] = E[e_{I_A}(X)] = \int_X e_{I_A}(x)f_X(x)dx. \quad (7.30)$$

Now

$$e_{I_A}(x) = \int_{Y_A} I_A(x, y)f_{Y \mid X}(y \mid x)dy$$

$$= \int_{A_\bar{x}} f_{Y \mid X}(y \mid x)dy$$

where $A_\bar{x} = \{y \in Y_X \mid (x, y) \in A\}. \quad (7.31)$

Using (7.30) in (7.31) yields

$$P[A] = \int_X \int_{A_\bar{x}} f_{Y \mid X}(y \mid x)f_X(x)dydx. \quad (7.32)$$
Chapter 7. Conditional distributions

which means that the joint pdf is \( f_{Y|X}(y \mid x)f_X(x) \).

What if \( X \) is discrete and \( Y \) is continuous, or vice versa? The joint formula (7.28) still works, but we have to interpret what the \( f(x, y) \) is. It is not a pmf or a pdf, but has elements of both. As long as we remember to sum over the discrete parts and integrate over the continuous parts, there should be no problems.

7.4 The marginal distribution of \( Y \)

There is no special trick in obtaining the marginal of \( Y \) given the conditional \( Y \mid X = x \) and the marginal of \( X \); just find the joint and integrate out \( x \). Thus

\[
 f_Y(y) = \int_{X_y} f_{Y|X}(y \mid x)f_X(x) \, dx. \tag{7.33}
\]

7.4.1 Example: Coins and the Beta-Binomial

In the coin example,

\[
 f_Y(y) = \int_0^1 \binom{n}{y} x^y (1-x)^{n-y} \, dx
 = \binom{n}{y} \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}
 = \frac{n!}{y!(n-y)!} \frac{y!(y-n)!}{(n+1)!}
 = \frac{1}{n+1}, \quad y = 0, 1, \ldots, n, \tag{7.34}
\]

which is the Discrete Uniform\( \{0, 1 \ldots, n\} \). Note that this distribution is not a binomial, and does not depend on \( x \) (it better not!). In fact, this \( Y \) is a special case of the following.

**Definition 7.** Suppose

\[
 Y \mid X = x \sim Binomial(n, x) \quad \text{and} \quad X \sim Beta(\alpha, \beta). \tag{7.35}
\]

Then the marginal distribution of \( Y \) is beta-binomial with parameters \( \alpha, \beta \) and \( n \), written

\[
 Y \sim Beta-Binomial(\alpha, \beta, n). \tag{7.36}
\]

When \( \alpha = \beta = 1 \), the \( X \) is uniform. Otherwise, as above, we can find the marginal pdf to be

\[
 f_Y(y) = \binom{n}{y} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}, \quad y = 0, 1, \ldots, n. \tag{7.37}
\]

7.4.2 Example: Simple normal linear model

For another example, take the linear model in (7.6) and (7.8),

\[
 Y \mid X = x \sim N(\alpha + \beta x, \sigma_c^2) \quad \text{and} \quad X \sim N(\mu_X, \sigma_X^2). \tag{7.38}
\]
7.4. The marginal distribution of $Y$

We could write out the joint pdf and integrate, but instead we will find the mgf, which we can do in steps because it is an expected value. That is, with $g(y) = e^{ty}$,

$$M_Y(t) = E[e^{tY}] = E[e^{g(X)}], \quad \text{where } e_g(x) = E[e^{tx} | X = x]. \quad (7.39)$$

We know the mgf of a normal, which $Y | X = x$ is, hence

$$e_g(x) = M_{N(\alpha_2 + \beta x, \sigma^2)}(t) = e^{(\alpha + \beta x)t + \frac{\sigma^2 t^2}{2}}. \quad (7.40)$$

The expected value of $e_g(X)$ can also be written as a normal mgf:

$$M_Y(t) = E[e^{g(X)}] = E[e^{(\alpha + \beta X)t + \frac{\sigma^2 t^2}{2}}] = e^{\alpha t + \frac{\sigma^2 t^2}{2} E[X]} = e^{\alpha t + \frac{\sigma^2 t^2}{2} M_{N(\mu_X, \sigma^2)}(\beta t)} = e^{\alpha t + \frac{\sigma^2 t^2}{2} e^{\beta \mu_X + \frac{\sigma^2 \beta^2}{2}}} = e^{(\alpha + \beta \mu_X t) + (\frac{\sigma^2 \beta^2}{2})} = \text{mgf of } N(\alpha + \beta \mu_X, \sigma^2 + \sigma_X^2 \beta^2). \quad (7.41)$$

That is, marginally,

$$Y \sim N(\alpha + \beta \mu_X, \sigma^2 + \sigma_X^2 \beta^2). \quad (7.42)$$

7.4.3 Marginal mean and variance

The marginal expected value of $Y$ is just the expected value of the conditional expected value:

$$E[Y] = E[e_Y(X)], \quad \text{where } e_Y(x) = E[Y | X = x]. \quad (7.43)$$

The marginal variance is not quite that simple. First, we need to define the conditional variance:

$$\nu_Y(x) \equiv \text{Var}[Y | X = x] = E[(Y - E[Y | X = x])^2 | X = x]. \quad (7.44)$$

It is the variance of the conditional distribution; just be sure to subtract the conditional mean before squaring. To find the marginal variance, we use $\text{Var}[Y] = e[Y^2] - E[Y]^2$.

Then

$$E[Y^2] = E[e_{Y^2}(X)], \quad \text{where } e_{Y^2}(x) = E[Y^2 | X = x]. \quad (7.45)$$

The conditional variance works like any variance, i.e.,

$$\text{Var}[Y | X = x] = E[Y^2 | X = x] - E[Y | X = x]^2, \quad (7.46)$$

hence

$$e_{Y^2}(x) = E[Y^2 | X = x] = \nu_Y(x) + e_Y(x)^2. \quad (7.47)$$

Thus

$$\text{Var}[Y] = E[Y^2] - E[Y]^2 = E[e_{Y^2}(X)] - (E[e_Y(X)])^2 = E[\nu_Y(X)] + E[e_Y(X)^2] - (E[e_Y(X)])^2 = E[\nu_Y(X)] + \text{Var}[e_Y(X)]. \quad (7.48)$$

To summarize:
The unconditional expected value is the expected value of the conditional expected value.

The unconditional variance is the expected value of the conditional variance plus the variance of the conditional expected value.

(The second sentence is very analogous to what happens in regression, where the total sum-of-squares equals the regression sum-of-squares plus the residual sum of squares.)

These are handy results. For example, in the beta-binomial (Definition 7), finding the mean and variance using the pdf (7.37) can be challenging. But using the conditional approach is much easier. Because conditionally \( Y \) is Binomial \((n, x)\),

\[
e_Y(x) = nx \quad \text{and} \quad v_Y(x) = nx(1 - x). \tag{7.49}
\]

Then because \( X \sim Beta(\alpha, \beta) \),

\[
E[Y] = nE[X] = n \frac{\alpha}{\alpha + \beta}, \tag{7.50}
\]

and

\[
Var[Y] = E[v_Y(X)] + Var[e_Y(X)] \\
= E[nX(1 - X)] + Var[nX] \\
= n \frac{\alpha \beta}{(\alpha + \beta)(\alpha + \beta + 1)} + n^2 \frac{\alpha^2}{(\alpha + \beta)^2(\alpha + \beta + 1)} \\
= \frac{n\alpha \beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{7.51}
\]

These expressions give some insight into the beta-binomial. Like the binomial, the beta-binomial counts the number of successes in \( n \) trials, and has expected value \( np \) for \( p = \alpha / (\alpha + \beta) \). Consider the variances of a binomial and beta-binomial:

\[
Var[Binomial] = np(1 - p) \quad \text{and} \quad Var[Beta-Binomial] = np(1 - p) \frac{\alpha + \beta + n}{\alpha + \beta + 1}. \tag{7.52}
\]

Thus the beta-binomial has a larger variance, so it can be used to model situations in which the data are more disperse than the binomial, e.g., if the \( n \) trials are \( n \) offspring in the same litter, and success is survival. The larger \( \alpha + \beta \), the closer the beta-binomial is to the binomial.

You might wish to check the mean and variance in the normal example 7.4.2.

The same procedure works for vectors:

\[
E[Y] = E[e_Y(X)] \quad \text{where} \quad e_Y(x) = E[Y | X = x], \tag{7.53}
\]

and

\[
Cov[Y] = E[v_Y(X)] + Cov[e_Y(X)], \quad \text{where} \quad v_Y(x) = Cov[Y | X = x]. \tag{7.54}
\]

In particular, considering a single covariance,

\[
Cov[Y_1, Y_2] = E[c_{Y_1,Y_2}(X)] + Cov[e_{Y_1}(X), e_{Y_2}(X)], \quad \text{where} \quad c_{Y_1,Y_2}(x) = Cov[Y_1, Y_2 | X = x]. \tag{7.55}
\]
7.4. The marginal distribution of $Y$

7.4.4 Example: Fruit flies, continued

Recall the fruit fly Example 1.1.1 from Chapter 1. There are $n = 10$ fathers, and each father has two offspring. Each father’s genotype is either $(TL, TL), (TL, CU),$ or $(CU, CU)$. We can represent the genotype by counting the number of $CU$’s, letting $X_i$ be the number (either 0, 1, or 2) of $CU$’s for the $i^{th}$ father. The mothers are all $(TL, TL)$, so each offspring receives a $TL$ from the mother, and randomly one of the alleles from the father. Let $Y_{ij}$ be the indicator of whether the $j^{th}$ offspring of father $i$ receives a $CU$ from the father. That is,

$$Y_{ij} = \begin{cases} 
0 & \text{if Offspring } j \text{ from father } i \text{ is } (TL, TL) \\
1 & \text{if Offspring } j \text{ from father } i \text{ is } (TL, CU).
\end{cases} \quad (7.56)$$

Then each “family” has three random variables, $(X_i, Y_{i1}, Y_{i2})$. We will assume that these triples are independent, in fact,

$$(X_1, Y_{11}, Y_{12}), \ldots, (X_n, Y_{n1}, Y_{n2}) \text{ are iid.} \quad (7.57)$$

To specify the distribution of each triple, we use the assumptions mentioned in the original Example 1.1.1. The proportion of $CU$’s in the population is $\theta$, and the assumptions imply that the number of $CU$’s for father $i$ is

$$X_i \sim \text{Binomial}(2, \theta), \quad (7.58)$$

because each father in effect randomly chooses two alleles from the population. Next, we specify the conditional distribution of the offspring given the father, i.e., $(Y_{i1}, Y_{i2}) \mid X_i = x_i)$. If $x_i = 0$, then the father is $(TL, TL)$, so the offspring will all receive a $TL$ from the father:

$$P[(Y_{i1}, Y_{i2}) = (0, 0) \mid X_i = 0] = 1. \quad (7.59)$$

Similarly, if $x_i = 2$, the father is $(CU, CU)$, so the offspring will all receive a $TL$ from the father:

$$P[(Y_{i1}, Y_{i2}) = (1, 1) \mid X_i = 2] = 1. \quad (7.60)$$

Finally, if $x_i = 1$, the father is $(TL, CU)$, which means each offspring has a 50-50 chance of receiving a $CU$ from the father:

$$P[(Y_{i1}, Y_{i2}) = (y_1, y_2) \mid X_i = 1] = \frac{1}{4} \text{ for } (y_1, y_2) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}. \quad (7.61)$$

This conditional distribution can be written more compactly by noting that $x_i/2$ is the chance that an offspring receives a $CU$, so that

$$(Y_{i1}, Y_{i2}) \mid X_i = x_i \sim \text{ iid Bernoulli}(x_i/2), \quad (7.62)$$

which using (7.60) yields the conditional pmf

$$f_{Y \mid X}(y_1, y_2 \mid x_i) = \left(\frac{x_i}{2}\right)^{y_1+y_2} \left(1 - \frac{x_i}{2}\right)^{2-y_1-y_2}. \quad (7.63)$$

The goal of the experiment is to estimate $\theta$, but without knowing the $X_i$’s. Thus the estimation has to be based on just the $Y_{ij}$’s. The marginal means are easy to find:

$$E[Y_{ij}] = E[e_{Y_{ij}}(X_i)], \text{ where } e_{Y_{ij}}(x_i) = E[Y_{ij} \mid X_i = x_i] = \frac{x_i}{2}, \quad (7.64)$$
Chapter 7. Conditional distributions

because conditionally \( Y_{ij} \) is Bernoulli. Then \( E[X_i] = 2\theta \), hence

\[
E[Y_{ij}] = \frac{2\theta}{2} = \theta. \tag{7.65}
\]

Nice! Then an obvious estimator of \( \theta \) is the sample mean of all the \( Y_{ij} \)'s, of which there are \( 2n = 20 \):

\[
\hat{\theta} = \frac{\sum_{i=1}^{2} \sum_{j=1}^{2} Y_{ij}}{2n}. \tag{7.66}
\]

The data are in Table 1.1. To find the estimate, we just count the number of CU’s, which is 8, hence the estimate of \( \theta \) is 0.4.

What is the variance of the estimate? Are \( Y_{i1} \) and \( Y_{i2} \) unconditionally independent? What is \( \text{Var}[Y_{ij}] \)? What is the marginal pmf of \( (Y_{i1}, Y_{i2}) \)?

## 7.5 The conditional from the joint

Often one has a joint distribution, but is primarily interested in the conditional, e.g., many experiments involve collecting health data from a population, and interest centers on the conditional distribution of certain outcomes, such as longevity, conditional on other variables, such as sex, age, cholesterol level, activity level, etc. In Bayesian inference, one can find the joint distribution of the data and the parameter, and from that needs to find the conditional distribution of the parameter given the data. Measure theory guarantees that for any joint distribution of \( (X, Y) \), there exists a conditional distribution of \( Y \mid X = x \) for each \( x \in X \). It may not be unique, but any conditional distribution that combines with the marginal of \( X \) to yield the original joint distribution is a valid conditional distribution. If densities exists, then \( f_{Y\mid X}(y \mid x) \) is a valid conditional density if (7.28) holds:

\[
f(x, y) = f_{Y\mid X}(y \mid x)f_X(x) \tag{7.67}
\]

for all \((x, y) \in \mathcal{W}\). Thus, given a joint density \( f \), one can integrate out \( y \) to obtain the marginal \( f_X \), then define the conditional by

\[
f_{Y\mid X}(y \mid x) = \frac{f(x, y)}{f_X(x)} = \frac{\text{Joint}}{\text{Marginal}} \quad \text{if} \quad f_X(x) > 0. \tag{7.68}
\]

It does not matter how the conditional density is defined when \( f_X(x) = 0 \), as long as it is a density on \( Y \), because in reconstructing the joint \( f \) in (7.67), it is multiplied by 0. Also, the conditional of \( X \) given \( Y = y \) is the ratio of the joint to the marginal of \( Y \).

This formula works for pdf’s, pmf’s, and the mixed kind.

### 7.5.1 Example: Coin

In the coin Example 2.4.3, the joint density of \( (X, Y) \) is

\[
f(x, y) = \binom{n}{y} x^y (1 - x)^{n-y}, \tag{7.69}
\]

and the marginal distribution of \( Y \), the number of heads, is, as in (7.34),

\[
f_Y(y) = \frac{1}{n+1}, y = 0, \ldots, n. \tag{7.70}
\]
Thus the conditional posterior distribution of $X$, the chance of heads, given $Y$, the number of heads, is

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)} = \frac{(n+1)!}{y!(n-y)!} \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} x^y(1-x)^{n-y} = \text{Beta}(y+1, n-y+1). \quad (7.71)$$

For example, if the experiment yields $Y = 3$ heads, then one’s guess of what the probability of heads is for this particular coin is described by the $	ext{Beta}(4, 8)$ distribution. So, for example, the best guess could be the posterior mean,

$$E[X \mid Y = 3] = E[\text{Beta}(4, 8)] = \frac{4}{4+8} = \frac{1}{3}. \quad (7.72)$$

Note that this is not the sample proportion of heads, 0.3, although it is close. The posterior mode and posterior median are also reasonable point estimates. A more informative quantity might be a probability interval:

$$P[0.1093 < X < 0.6097 \mid Y = 3] = 95\%. \quad (7.73)$$

So there is a 95% chance that the chance of heads is somewhere between 0.11 and 0.61. (These numbers were found using the $qbeta$ function in R.) It is not a very tight interval, but there is not much information in just ten flips.

### 7.5.2 Bivariate normal

If one can recognize the form of a particular density for $Y$ within a joint density, then it becomes unnecessary to explicitly find the marginal of $X$, then divide the joint by the marginal. For example, the $N(\mu, \sigma^2)$ density can be written

$$\phi(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left( \frac{(z-\mu)^2}{\sigma^2} \right)} = c(\mu, \sigma^2) e^{-\frac{1}{2} \left( \frac{z^2}{\sigma^2} \right)}. \quad (7.74)$$

That is, we factor the pdf into a constant we do not care about at the moment, that depends on the fixed parameters, and the important component containing all the $z$-action.

Now consider the bivariate normal,

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\mu, \Sigma), \text{ where } \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}. \quad (7.75)$$

Assuming $\Sigma$ is invertible, the joint pdf is (as in (6.53))

$$f(x, y) = \frac{1}{2\pi} \frac{1}{\sqrt{|\Sigma|}} e^{-\frac{1}{2} \left( (x,y)' \Sigma^{-1} (x,y)' - \mu \right)}. \quad (7.76)$$
The conditional pdf of $Y \mid X = x$ can be written

$$f_{Y \mid X}(y \mid x) = \frac{f(x, y)}{f_X(x)} = c(x, \mu, \Sigma) \times ??,$$  (7.77)

where ?? is a function of $y$ that, we hope, is the core of a recognizable density. The $c$ contains the $f_X(x)$, plus other factors, so we do not need to explicitly find the marginal. To try to isolate the parts depending on $y$, take part of the exponent:

$$(x, y)\Sigma^{-1}(x, y) = (x, y)\Sigma^{-1}(x, y)' - 2(x, y)\Sigma^{-1} \mu + \mu'\Sigma^{-1} \mu.$$  (7.78)

Recall that

$$\Sigma^{-1} = \frac{1}{|\Sigma|} \begin{pmatrix} \sigma_Y^2 & -\sigma_{XY} \\ -\sigma_{XY} & \sigma_X^2 \end{pmatrix}.$$  (7.79)

Now the expression in (7.78), considered as a function of $y$, is quadratic, i.e., $a + by + cy^2$, where $a$, $b$, and $c$ depend on $\mu, \Sigma$ and $x$. We do not care about the $a$, but the linear and quadratic terms are important. Some algebra yields

$$(x, y)\Sigma^{-1}(x, y)' = -2\frac{x\sigma_{XY}}{|\Sigma|} y + \frac{\sigma_X^2}{|\Sigma|} y^2 + \cdots,$$  and

$$-2(x, y)\Sigma^{-1} \mu = -2 \frac{-\mu x \sigma_{XY} + \mu y \sigma_X^2}{|\Sigma|} y + \cdots.$$  (7.80)

Then

$$f_{Y \mid X}(y \mid x) = c(x, \mu, \sigma^2) e^Q,$$  (7.81)

where

$$Q = -\frac{1}{2} \left( -2 \frac{x \sigma_{XY}}{|\Sigma|} y + \frac{\sigma_X^2}{|\Sigma|} y^2 - 2 \frac{-\mu x \sigma_{XY} + \mu y \sigma_X^2}{|\Sigma|} y \right)$$

$$= \frac{\sigma_{XY} (x - \mu_X) + \mu y \sigma_X^2}{|\Sigma|} y - \frac{1}{2} \frac{\sigma_X^2}{|\Sigma|} y^2.$$  (7.82)

Now compare the $Q$ in (7.82) to the exponent in (7.74). We see that $f_{Y \mid X}(y \mid x)$ is of the same form, with

$$\frac{1}{\sigma^2} = \frac{\sigma_X^2}{|\Sigma|} \quad \text{and} \quad \frac{\mu}{\sigma^2} = \frac{(x - \mu_X) \sigma_{XY} + \mu y \sigma_X^2}{|\Sigma|}.$$  (7.83)

Thus we just need to solve for $\mu$ and $\sigma^2$, and will then have that $Y \mid X = x \sim N(\mu, \sigma^2)$. To proceed,

$$\sigma^2 = \frac{|\Sigma|}{\sigma_X^2} = \frac{\sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2} = \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2},$$  (7.84)

and

$$\mu = \frac{(x - \mu_X) \sigma_{XY} + \mu y \sigma_X^2}{\sigma_X^2} = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2} (x - \mu_X).$$  (7.85)

That is,

$$Y \mid X = x \sim N \left( \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2} (x - \mu_X), \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2} \right).$$  (7.86)
Because we know the normal pdf, we could work backwards to find the \( c \) in (7.77), but there is no need to do that.

This is a normal linear model, as in (7.6), where

\[
Y | X = x \sim N(\alpha + \beta x, \sigma^2_x) .
\]  

(7.87)

We can make the identifications:

\[
\beta = \frac{\sigma_{XY}}{\sigma_X^2}, \quad \alpha = \mu_Y - \beta \mu_X, \quad \text{and} \quad \sigma^2_e = \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2} .
\]

(7.88)

These equations should be familiar from linear regression.

### 7.6 Bayes Theorem: Reversing the conditionals

We have already essentially derived Bayes Theorem, but it is important enough to deserve its own section. The theorem takes the conditional density of \( Y \) given \( X \) and the marginal distribution of \( X \), and produces the conditional density of \( X \) given \( Y \).

It uses the formula \( \text{conditional} = \text{joint} / \text{marginal} \), where the marginal is found by integrating out the \( y \) from the joint.

**Theorem 3. Bayes Theorem** Suppose \( Y | X = x \) has density \( f_{Y|X}(y \mid x) \) and \( X \) has marginal density \( f_X(x) \). Then,

\[
f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x)f_X(x)}{\int_{X} f_{Y|X}(y \mid x)f_X(x)dx} .
\]

(7.89)

The integral will be a summation if \( X \) is discrete.

**Proof** With \( f(x, y) \) being the joint density,

\[
f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)} = \frac{f_{Y|X}(y \mid x)f_X(x)}{\int_X f_{Y|X}(y \mid x)f_X(x)dx} = \frac{f_{Y|X}(y \mid x)f_X(x)}{\int_X f_{Y|X}(y \mid x)f_X(x)dx} . \quad \square
\]

(7.90)

Bayes Theorem is often used with sets. Let \( A \subset X \) and \( B_1, \ldots, B_K \) be a partition of \( X \), i.e.,

\[
B_i \cap B_j = \emptyset \text{ for } i \neq j, \quad \text{and} \quad \bigcup_{k=1}^K B_k = X .
\]

(7.91)

Then

\[
P[B_k \mid A] = \frac{P[A \mid B_k]P[B_k]}{\sum_{i=1}^K P[A \mid B_i]P[B_i]} .
\]

(7.92)
7.6.1 Example: AIDS virus

A common illustration of Bayes Theorem involves testing for some medical condition, e.g., a blood test for the AIDS virus. Suppose the test is 99% accurate. If a random person’s test is positive, does that mean the person is 99% sure of having the virus?

Let \( A_+ = \text{“test is positive”} \), \( A_- = \text{“test is negative”} \), \( B_v = \text{“person has the virus”} \) and \( B_n = \text{“person does not have the virus.”} \) Then we know the conditionals

\[
P[A_+ | B_v] = 0.99 \quad \text{and} \quad P[A_- | B_n] = 0.99,
\]

but they are not of interest. We want to know the reverse conditional, \( P[B_v | A_+] \), the chance of having the virus given the test is positive. There is no way to figure this probability out without the marginal of \( B \), that is, the marginal chance a random person has the virus. Let us say that \( P[B_v] = 1/10,000 \). Now we can use Bayes Theorem (7.92)

\[
P[B_v | A_] = \frac{P[A_+ | B_v]P[B_v]}{P[A_+ | B_v]P[B_v] + P[A_+ | B_n]P[B_n]}
\]

\[
= \frac{0.99 \times \frac{1}{10000}}{0.99 \times \frac{1}{10000} + 0.01 \times \frac{9999}{10000}}
\]

\[
\approx 0.0098.
\]

Thus the chance of having the virus, given the test is positive, is only about 1/100. That is lower than one might expect, but it is substantially higher than the overall chance of 1/10000. (This example is a bit simplistic in that random people do not take the test, but more likely people who think they may be at risk.)

7.6.2 Example: Beta posterior for the binomial

The coin Example 7.5.1 can be generalized by using a beta in place of the uniform. In a Bayesian framework, we suppose the probability of success, \( \theta \), has a prior distribution \( \text{Beta}(\alpha, \beta) \), and \( Y | \theta = \theta \) is \( \text{Binomial}(n, \theta) \). (So now \( x \) has become \( \theta \).) The prior is supposed to represent knowledge or belief about what the \( \theta \) is before seeing the data \( Y \). To find the posterior, or what we are to think after seeing the data \( Y = y \), we need the conditional distribution of \( \theta \) given \( Y = y \).

The joint density is

\[
f(\theta, y) = f_Y(y | \theta)f_\theta(\theta) = \binom{n}{y}\theta^y(1 - \theta)^{n-y}\beta(\alpha, \beta)\beta^{\alpha-1}(1 - \theta)^{\beta-1}
\]

\[
= c(y, \alpha, \beta)\theta^{y+\alpha-1}(1 - \theta)^{n-y\beta-1}.
\]

Because we are interested in the pdf of \( \theta \), we put everything not depending on \( \theta \) in the constant. But the part that does depend on \( \theta \) is the meat of a \( \text{Beta}(\alpha + y, \beta + n - y) \) density, hence that is the posterior:

\[
\theta | Y = y \sim \text{Beta}(\alpha + y, \beta + n - y).
\]

7.7 Conditionals and independence

If \( X \) and \( Y \) are independent, then it makes sense that the distribution of one does not depend on the value of the other, which is true.
Lemma 9. The random vectors \( X \) and \( Y \) are independent if and only if the conditional distribution\(^1\) of \( Y \) given \( X = x \) does not depend on \( x \).

When there are densities, this result follows directly:

\[
\text{Independence} \quad \Rightarrow \quad f(x, y) = f_X(x)f_Y(y) \quad \text{and} \quad \mathcal{W} = \mathcal{X} \times \mathcal{Y}
\]

\[
\Rightarrow \quad f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)} = f_Y(y) \quad \text{and} \quad \mathcal{Y}_x = \mathcal{Y},
\]

which does not depend on \( x \). The other way, if the conditional distribution of \( Y \) given \( X = x \) does not depend on \( x \), then \( f_{Y|X}(y \mid x) = f_Y(y) \) and \( \mathcal{Y}_x = \mathcal{Y} \), hence

\[
f(x, y) = f_{Y|X}(y \mid x)f_X(x) = f_Y(y)f_X(x) \quad \text{and} \quad \mathcal{W} = \mathcal{X} \times \mathcal{Y}.
\]

7.7.1 Example: Independence of residuals and \( X \)

Suppose \((X, Y)\)' is bivariate normal as in (7.75),

\[
\begin{pmatrix}
X \\
Y
\end{pmatrix} \sim N\left(\begin{pmatrix}
\mu_X \\
\mu_Y
\end{pmatrix}, \begin{pmatrix}
\sigma^2_X & \sigma_{XY} \\
\sigma_{XY} & \sigma^2_Y
\end{pmatrix}\right).
\]

(7.99)

We then have that

\[
Y \mid X = x \sim N(\alpha + \beta x, \sigma^2_e),
\]

(7.100)

where the parameters are given in (7.88). The residual is \( Y - \alpha - \beta X \). What is its conditional distribution? First, for fixed \( x \),

\[
Y - \alpha - \beta X \mid X = x = \mathcal{D} Y - \alpha - \beta x \mid X = x.
\]

(7.101)

This equation means that when we are conditioning on \( X = x \), the conditional distribution stays the same if we fix \( X = x \), which follows from the original definition of conditional expected value in (7.15). We know that subtracting the mean from a normal leaves a normal with mean 0 and the same variance, hence

\[
Y - \alpha - \beta x \mid X = x \sim N(0, \sigma^2_e).
\]

(7.102)

But the right-hand side has no \( x \), hence \( Y - \alpha - \beta X \) is independent of \( X \), and has marginal distribution \( N(0, \sigma^2_e) \).

\(^1\)Measure-theoretically, we technically mean that “a version of the conditional distribution ...”
Chapter 8

The Multivariate Normal Distribution

Almost all data are multivariate, that is, entail more than one variable. There are two general-purpose multivariate models: the multivariate normal for continuous data, and the multinomial for categorical data. There are many specialized multivariate distributions, but these two are the only ones that are used in all areas of statistics. We have seen the bivariate normal in Example 6.2.4, and the multinomial is introduced in Example 3.4.3. This chapter focuses on the multivariate normal.

8.1 Definition

There are not very many commonly used multivariate distributions to model data. The multivariate normal is by far the most common, at least for continuous data. Which is not to say that all data are distributed normally, nor that all techniques assume such. Rather, one either assumes normality, or makes few assumptions at all and relies on asymptotic results. Some of the nice properties of the multivariate normal:

- It is completely determined by the means, variances, and covariances.
- Elements are independent if and only if they are uncorrelated.
- Marginals of multivariate normals are multivariate normal.
- An affine transformation of a multivariate normal is multivariate normal.
- Conditionals of a multivariate normal are multivariate normal.
- The sample mean is independent of the sample covariance matrix.

The multivariate normal arises from iid standard normals, that is, iid $N(0,1)$'s. Suppose $Z = (Z_1, \ldots, Z_M)'$ is an $M \times 1$ vector of iid $N(0,1)$'s. Because they are independent, all the covariances are zero, so that

$$E[Z] = 0_M \quad \text{and} \quad \text{Cov}[Z] = I_M.$$ (8.1)

A general multivariate normal distribution can have any (legitimate) mean and covariance, achieved through the use of affine transformations. Here is the definition.
Chapter 8. The Multivariate Normal Distribution

Definition 8. Let the \( p \times 1 \) vector \( \mathbf{Y} \) be defined by

\[
\mathbf{Y} = \mu + BZ,
\]

where \( B \) is \( p \times M \), \( \mu \) is \( p \times M \), and \( Z \) is an \( M \times 1 \) vector of iid standard normals. Then \( \mathbf{Y} \) is multivariate normal with mean \( \mu \) and covariance matrix \( \Sigma \equiv BB' \), written

\[
\mathbf{Y} \sim N(\mu, \Sigma).
\]

From (8.1) and (8.2), the usual affine transformation results from Section 3.5 show that \( \mathbb{E}[\mathbf{Y}] = \mu \) and \( \text{Cov}[\mathbf{Y}] = BB' = \Sigma \). The definition goes further, implying that the distribution depends on \( B \) only through the \( BB' \). For example, consider the two matrices

\[
B_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 3/\sqrt{5} & 1/\sqrt{5} \\ 0 & \sqrt{5} \end{pmatrix},
\]

so that

\[
B_1B_1' = B_2B_2' = \begin{pmatrix} 2 & 1 \\ 1 & 5 \end{pmatrix} = \Sigma.
\]

Thus the definition says that both

\[
Y_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} Z_1 \\ uZ_2 \\ uZ_3 \end{pmatrix}, \quad Y_2 = \begin{pmatrix} 3/\sqrt{5} & 1/\sqrt{5} \\ 0 & \sqrt{5} \end{pmatrix} \begin{pmatrix} Z_1 \\ uZ_2 \end{pmatrix}
\]

are \( N(0_2, \Sigma) \). They clearly have the same mean and covariance matrix, but it is not obvious they have the exact same distribution, especially as they depend on different numbers of \( Z_i \)'s. To see the distributions are the same, we have to look at the mgf. We have already found the mgf for the bivariate \((p = 2)\) normal in (6.57), and the proof here is the same. The answer is

\[
M_Y(t) = \exp(\mu' t + \frac{1}{2} t' \Sigma t), \quad t \in \mathbb{R}^p.
\]

Thus the distribution of \( Y \) does depend on just \( \mu \) and \( \Sigma \), so that because \( Y_1 \) and \( Y_2 \) have the same \( B_iB_i' \) (and same mean), they have the same distribution.

Can \( \mu \) and \( \Sigma \) be anything, or are there restrictions? Any \( \mu \) is possible, since \( \mu = a \) and \( a \) can be anything in \( \mathbb{R}^p \). The covariance matrix \( \Sigma \) can be \( BB' \) for any \( p \times M \) matrix \( B \). Note that \( M \) is arbitrary, too. Clearly, \( \Sigma \) must be symmetric, but we already knew that. It must also be nonnegative definite, which we define now.

Definition 9. A symmetric \( p \times p \) matrix \( B \) is nonnegative definite if

\[
\mathbf{c}'B\mathbf{c} \geq 0 \text{ for all } p \times 1 \text{ vectors } \mathbf{c}.
\]

Also, \( B \) is positive definite if

\[
\mathbf{c}'B\mathbf{c} > 0 \text{ for all } p \times 1 \text{ vectors } \mathbf{c} \neq 0.
\]
8.1. Definition

Note that \( \mathbf{c}' \mathbf{B} \mathbf{c} = \| \mathbf{c}' \mathbf{B} \|^2 \geq 0 \), which means that \( \Sigma \) must be nonnegative definite. But from (3.103),

\[
\mathbf{c}' \Sigma \mathbf{c} = \operatorname{Cov}(\mathbf{c}' \mathbf{Y}) = \operatorname{Var}(\mathbf{c}' \mathbf{Y}) \geq 0,
\]

(8.10)
because all variances are nonnegative. That is, any covariance matrix has to be nonnegative definite, not just multivariate normal ones.

So we know that \( \Sigma \) must be symmetric and nonnegative definite. Are there any other restrictions, or for any symmetric nonnegative definite matrix is there a corresponding \( \mathbf{B} \)? In fact, there are potentially many such square roots \( \mathbf{B} \). A nice one is the symmetric square root available because \( \Sigma \) is symmetric. See the next subsection.

We conclude that the multivariate normal distribution is defined for and \((\mu, \Sigma)\), where \( \mu \in \mathbb{R} \) and \( \Sigma \) is symmetric and positive definite, i.e., it can be any valid covariance matrix.

8.1.1 The spectral decomposition

To derive a symmetric square root, as well as perform other useful tasks, we need the following decomposition.

**Theorem 4. The Spectral Decomposition Theorem for Symmetric Matrices.** If \( \Omega \) is a symmetric \( p \times p \) matrix, then there exists a \( p \times p \) orthogonal (6.58) matrix \( \mathbf{\Gamma} \) and a \( p \times p \) diagonal matrix \( \mathbf{\Lambda} \) with diagonals \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \) such that

\[
\Omega = \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}'.
\]

(8.11)

Write \( \mathbf{\Gamma} = (\gamma_1, \ldots, \gamma_p) \), so that the \( \gamma_i \)'s are the columns of \( \mathbf{\Gamma} \). Then

\[
\Omega = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i'.
\]

(8.12)

Because the \( \gamma_i \)'s are orthogonal, and each have length 1,

\[
\Omega \gamma_i = \lambda_i \gamma_i,
\]

(8.13)

which means \( \gamma_i \) is an eigenvector of \( \Omega \), with corresponding eigenvalue \( \lambda_i \). The eigenvalues are unique, but the eigenvectors will not all be unless the eigenvectors are distinct. For example, the identity matrix \( \mathbf{I}_p \) has all eigenvalues 1, and any vector \( \gamma \) with length 1 is an eigenvector.

Here are some handy facts about symmetric \( \Omega \) and its spectral decomposition.

- \( \Omega \) is positive definite if and only if all \( \lambda_i \)'s are positive, and nonnegative definite if and only if all \( \lambda_i \)'s are nonnegative.
- The trace and determinant are, respectively,

\[
\operatorname{trace}(\Omega) = \sum \lambda_i \quad \text{and} \quad |\Omega| = \prod \lambda_i.
\]

(8.14)

(The trace of a square matrix is the sum of its diagonals.)

- \( \Omega \) is invertible if and only if its eigenvalues are nonzero, in which case its inverse is

\[
\Omega^{-1} = \mathbf{\Gamma} \mathbf{\Lambda}^{-1} \mathbf{\Gamma}'.
\]

(8.15)

Thus the inverse has the same eigenvectors, and eigenvalues \( 1/\lambda_i \).
• If $\Omega$ is nonnegative definite, then with $\Lambda^{1/2}$ being the diagonal matrix with diagonal elements $\sqrt{\lambda_i}$,

$$\Omega^{1/2} = \Gamma \Lambda^{1/2} \Gamma'$$

(8.16)

is a symmetric square root of $\Omega$, that is, it is symmetric and $\Omega^{1/2} \Omega^{1/2} = \Omega$.

The last item was used in the previous section to guarantee that any covariance matrix has a square root.

### 8.2 Some properties of the multivariate normal

We will prove the next few properties using the representation (8.2). They can also be easily shown using the mgf.

#### 8.2.1 Affine transformations

Affine transformations of multivariate normals are also multivariate normal, because any affine transformation of a multivariate normal vector is an affine transformation of an affine transformation of a standard normal vector, and an affine transformation of an affine transformation is also an affine transformation. That is, suppose $\mathbf{Y} \sim N_p(\mu, \Sigma)$, and $\mathbf{W} = \mathbf{c} + D\mathbf{Y}$ for $q \times p$ matrix $D$ and $q \times 1$ vector $\mathbf{c}$. Then we know that for some $B$ with $BB' = \Sigma$, $\mathbf{Y} = \mu + B \mathbf{Z}$, where $\mathbf{Z}$ is a vector of iid standard normals. Hence

$$\mathbf{W} = \mathbf{c} + D\mathbf{Y} = \mathbf{c} + D(\mu + B \mathbf{Z}) = \mathbf{c} + D\mu + DB\mathbf{Z}.$$  

(8.17)

Then by Definition 8,

$$\mathbf{W} \sim N(\mathbf{c} + D\mu, DBB'D').$$

(8.18)

Of course, the mean and covariance result we already knew.

#### 8.2.2 Marginals

Because marginals are special cases of affine transformations, marginals of multivariate normals are also multivariate normal. One needs just to pick off the appropriate means and covariances. So if $\mathbf{Y} = (Y_1, \ldots, Y_5)'$ is $N_5(\mu, \Sigma)$, and $\mathbf{W} = (Y_2, Y_5)'$, then

$$\mathbf{W} \sim N_2\left(\begin{pmatrix} \mu_2 \\ \mu_5 \end{pmatrix}, \begin{pmatrix} \sigma_{22} & \sigma_{25} \\ \sigma_{52} & \sigma_{55} \end{pmatrix}\right).$$

(8.19)

Here, $\sigma_{ij}$ is the $ij^{th}$ element of $\Sigma$, so that $\sigma_{ii} = \sigma_i^2$.

#### 8.2.3 Independence

In Section 4.3, we showed that independence of two random variables means that their covariance is 0, but that a covariance of 0 does not imply independence. But, with multivariate normals, it does. That is, if $\mathbf{X}$ is a multivariate normal vector, and $\text{Cov}(X_j, X_k) = 0$, then $X_j$ and $X_k$ are independent. The next theorem generalizes this independence to sets of variables.
Theorem 5. Suppose
\[ X = \begin{pmatrix} Y \\ W \end{pmatrix} \] (8.20)

is multivariate normal, where \( Y = (Y_1, \ldots, Y_K)' \) and \( W = (W_1, \ldots, W_L)' \). If \( \text{Cov}(Y_k, W_l) = 0 \) for all \( k, l \), then \( Y \) and \( W \) are independent.

Proof. For simplicity, we will assume the mean of \( X \) is 0\(_{K+L} \). Because covariances between the \( Y_k \)'s and \( W_l \)'s are zero, \( \text{Cov}(X) = \begin{pmatrix} \text{Cov}(Y) & 0 \\ 0 & \text{Cov}(W) \end{pmatrix} \). (8.21)

(The 0's denote matrices of the appropriate size with all elements zero.) Each of those individual covariance matrices has a square root, hence there are matrices \( B, K \times K \), and \( C, L \times L \), such that
\[ \text{Cov}(Y) = BB' \quad \text{and} \quad \text{Cov}(W) = CC'. \] (8.22)

Thus
\[ \text{Cov}(X) = AA' \quad \text{where} \quad A = \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix}. \] (8.23)

Then by definition, we know that with \( Z \) being a \((K + L) \times 1\) vector of iid standard normals,
\[ \begin{pmatrix} Y \\ W \end{pmatrix} = X = AZ = \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} BZ_1 \\ CZ_2 \end{pmatrix}, \] (8.24)

where \( Z_1 \) and \( Z_2 \) are \( K \times 1 \) and \( L \times 1 \), respectively. But that means that
\[ Y = BZ_1 \quad \text{and} \quad W = CZ_2, \] (8.25)

and \( Z_1 \) is independent of \( Z_2 \), hence \( Y \) is independent of \( W \). \( \Box \)

Note that this result can be extended to partitions of \( X \) into more than two groups. Especially, if all the covariances between elements of \( X \) are 0, then the elements are mutually independent.

### 8.3 The pdf

The multivariate normal has a pdf only if the covariance is invertible (i.e., positive definite). In that case, its pdf is easy to find using the same procedure used to find the pdf of the bivariate normal in Section 6.2.4. Suppose \( Y \sim N(\mu, \Sigma) \) where \( \Sigma \) is a \( p \times p \) positive definite matrix. Let \( B \) be the symmetric square root of \( \Sigma \), which will also be positive definite. (Why?\(^1\)) Then
\[ Y = \mu + BZ, \quad \begin{pmatrix} Z \end{pmatrix} \sim N(0_p, I_p). \] (8.26)

We know the pdf of \( Z \) because it consists of iid \( N(0,1) \)'s:
\[ f_Z(z) = \prod \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2} = \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2} \|z\|^2}. \] (8.27)

\(^1\)Because the eigenvalues of \( \Sigma \) are positive, and those of \( B \) are the square roots, so are also positive.
The inverse of the transformation (8.26) is
\[ Z = B^{-1}(Y - \mu), \]  
(8.28)
hence the Jacobian is
\[ |B^{-1}| = 1/|B|. \]
Then
\[ f_Y(y) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|B|} e^{-\frac{1}{2} ||B^{-1}(y-\mu)||^2}. \]  
(8.29)
Because \( \Sigma = BB' = BB, |\Sigma| = |B|^2. \) Also
\[ ||B^{-1}(y-\mu)||^2 = (y-\mu)'(BB)^{-1}(y-\mu). \]  
(8.30)
Thus,
\[ f_Y(y) = \frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{|\Sigma|}} e^{-\frac{1}{2} (y-\mu)'\Sigma^{-1}(y-\mu)}. \]  
(8.31)

8.4 The sample mean and variance

Often one desires a confidence interval for the population mean. Specifically, suppose \( X_1, \ldots, X_n \) are iid \( N(\mu, \sigma^2) \). Then
\[ \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1), \]  
(8.32)
because \( \overline{X} \sim N(\mu, \sigma^2/n) \). Because \( P[-1.96 < N(0,1) < 1.96] = 0.95, \)
\[ P[-1.96 < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < 1.96] = 0.95, \]  
(8.33)
or, untangling the equations to get \( \mu \) in the middle,
\[ P[\overline{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96 \frac{\sigma}{\sqrt{n}}] = 0.95. \]  
(8.34)
Thus,
\[ (\overline{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96 \frac{\sigma}{\sqrt{n}}) \] is a 95% confidence interval for \( \mu, \)  
(8.35)
at least if \( \sigma \) is known. But what if \( \sigma \) is not known? Then you estimate it, which will change the distribution, that is,
\[ \frac{\overline{X} - \mu}{\hat{\sigma} / \sqrt{n}} \sim ??. \]  
(8.36)
The sample variance for a sample \( x_1, \ldots, x_n \) is
\[ s^2 = \frac{\sum(x_i - \overline{x})^2}{n}, \]  
or is it \( s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} ? \)  
(8.37)
Rather than worry about that question now, we will find the joint distribution of
\[ (\overline{X}, U), \] where \( U = \sum(X_i - \overline{X})^2. \)  
(8.38)
To start, note that the deviations \( x_i - \bar{x} \) are linear functions of the \( x_i \)'s, as of course is \( \bar{x} \), and we know how to deal with linear combinations of normals. That is, letting \( \mathbf{X} = (X_1, \ldots, X_n)' \), because the elements are iid,

\[
\mathbf{X} \sim N(\mu_1n, \sigma^2 I_n),
\]

where \( 1_n \) is the \( n \times 1 \) vector of all 1's. The mean and deviations can be written

\[
\begin{pmatrix}
\bar{X} \\
X_1 - \bar{X} \\
\vdots \\
X_n - \bar{X}
\end{pmatrix} = \begin{pmatrix}
0 \\
1_n \\
\vdots \\
1_n
\end{pmatrix} \mathbf{X} - \begin{pmatrix}
-1 \\
-1_n \\
\vdots \\
-1_n
\end{pmatrix} \mathbf{X}.
\]

Now \( \mathbf{X} = I_n \mathbf{X} \) and \( \bar{X} = (1/n)1_n' \mathbf{X} \), hence

\[
\begin{pmatrix}
\bar{X} \\
X_1 - \bar{X} \\
\vdots \\
X_n - \bar{X}
\end{pmatrix} = \begin{pmatrix}
0 \\
1_n \\
\vdots \\
1_n
\end{pmatrix} \mathbf{X} - \begin{pmatrix}
-1/1_n \\
-1_n/1_n \\
\vdots \\
-1_n/1_n
\end{pmatrix} \mathbf{X}
\]

\[
= \begin{pmatrix}
1/1_n 1_n' \\
H_n
\end{pmatrix} \mathbf{X},
\]

where

\[
H_n = I_n - \frac{1}{n} 1_n 1_n'
\]

is the \( n \times n \) centering matrix. It is called the centering matrix because for any \( n \times 1 \) vector \( \mathbf{a} \), \( H_n \mathbf{a} \) subtracts the mean of the elements from each element, centering the values at 0. Note that if all the elements are the same, centering will set everything to 0, i.e.,

\[
H_n 1_n = 0_n.
\]

Also, if the mean of the elements already is 0, centering does nothing, which in particular means that \( H_n (H_n \mathbf{a}) = H_n \mathbf{a} \), or

\[
H_n H_n = H_n.
\]

Such a matrix is called idempotent. It is not difficult to verify (8.44) directly using the definition (8.42) and multiplying things out. In fact, \( I_n \) and \( (1/n)1_n 1_n' \) are also idempotent.

Back to the task. Equation (8.41) gives explicitly that the mean and deviations are a linear transformation of a multivariate normal, hence the vector is multivariate normal. The mean and covariance are

\[
E \left[ \begin{pmatrix}
\bar{X} \\
X_1 - \bar{X} \\
\vdots \\
X_n - \bar{X}
\end{pmatrix} \right] = \begin{pmatrix}
1/1_n 1_n' \\
H_n
\end{pmatrix} \mu_1n = \begin{pmatrix}
1/1_n 1_n' 1_n \\
H_n 1_n
\end{pmatrix} \mu = \begin{pmatrix}
\mu \\
0_n
\end{pmatrix}
\]

(8.45)
Chapter 8. The Multivariate Normal Distribution

and

\[
\begin{align*}
\text{Cov} \left[ \begin{pmatrix} \bar{X} \\ X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix} \right] &= \begin{pmatrix} \frac{1}{n} \mathbf{1}_n' \\ \mathbf{H}_n \end{pmatrix} \sigma^2 \mathbf{I}_n \begin{pmatrix} \frac{1}{n} \mathbf{1}_n' \\ \mathbf{H}_n \end{pmatrix}' \\
&= \sigma^2 \begin{pmatrix} \frac{1}{n} \mathbf{1}_n' \mathbf{1}_n & \frac{1}{n} \mathbf{1}_n' \mathbf{H}_n \\
\frac{1}{n} \mathbf{H}_n \mathbf{1}_n' & \mathbf{H}_n \mathbf{H}_n \end{pmatrix} \\
&= \sigma^2 \begin{pmatrix} \frac{1}{n} & 0' \\
0' & \mathbf{H}_n \end{pmatrix}. 
\end{align*}
\]

(8.46)

Look at the \(0_n\)'s: The covariances between \(\bar{X}\) and the deviations \(\mathbf{H}_nX\) are zero, hence with the multivariate normality means they are independent. Further, we can read off the distributions:

\[
\bar{X} \sim N(\mu, \frac{1}{n}\sigma^2) \quad \text{and} \quad \mathbf{H}_nX \sim N(0_n, \sigma^2\mathbf{H}_n).
\]

(8.47)

The first we already knew. But because \(\bar{X}\) and \(\mathbf{H}_nX\) are independent, and \(U = \|\mathbf{H}_nX\|^2\) is a function of just \(\mathbf{H}_nX\),

\[
\bar{X} \quad \text{and} \quad U = \sum(X_i - \bar{X})^2 \quad \text{are independent.}
\]

(8.48)

The next section will find the distribution of \(U\).

8.5 Chi-squares

In Section 5.1.2, we defined the chi-squared distribution on one degree of freedom. Here we look at the more general chi-square.

**Definition 10.** Suppose \(Z \sim N(\mu, \mathbf{I}_v)\). Then

\[
W = \|Z\|^2
\]

(8.49)

has the **noncentral chi-squared distribution** on \(v\) degrees of freedom with noncentrality \(\Delta = \|\mu\|^2\), written

\[
W \sim \chi^2_v(\Delta).
\]

(8.50)

If \(\Delta = 0\), then \(W\) has the (central) chi-squared distribution,

\[
W \sim \chi^2_v.
\]

(8.51)

This definition implies that if two \(\mu\)'s have the same length, then the corresponding \(W\)'s have the same distribution. That is, with \(v = 2\),

\[
\mu = (5, 0), \quad \mu = (3, 4), \quad \text{and} \quad \mu = (\frac{5}{\sqrt{2}}, \frac{5}{\sqrt{2}})
\]

(8.52)
all lead to the same $\Delta = 25$, hence the distributions of their $\|Z\|^2$’s will all be the same, $\chi^2(25)$. Is that a plausible fact? The key is that if $\Gamma$ is an orthogonal matrix, then $\|\Gamma Z\| = \|Z\|$. Thus $Z$ and $\Gamma Z$ would lead to the same chi-squared. Take $Z \sim N(\mu, I_\nu)$, and let $\Gamma$ be the orthogonal matrix such that

$$\Gamma \mu = \begin{pmatrix} \|\mu\| \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \sqrt{\Delta} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (8.53)$$

Any orthogonal matrix whose first row is $\mu/\|\mu\|$ will work. Then

$$\|Z\|^2 \equiv \|\Gamma Z\|^2, \quad (8.54)$$

and the latter clearly depends on just $\Delta$, which shows that the definition is fine.

The definition can be used to show that the sum of independent chi-squared is also chi-squared.

**Lemma 10.** If $W_1$ and $W_2$ are independent, with

$$W_1 \sim \chi^2_{\nu_1}(\Delta_1) \text{ and } W_2 \sim \chi^2_{\nu_2}(\Delta_2), \text{ then } W_1 + W_2 \sim \chi^2_{\nu_1+\nu_2}(\Delta_1 + \Delta_2). \quad (8.55)$$

**Proof.** Because $W_1$ and $W_2$ are independent, they can be represented by $W_1 = \|Z_1\|^2$ and $\|Z_2\|^2$, where $Z_1$ and $Z_2$ are independent, and

$$Z_1 \sim N(\mu_1, I_{\nu_1}) \text{ and } Z_2 \sim N(\mu_2, I_{\nu_2}), \text{ $\Delta_1 = \|\mu_1\|^2$ and $\Delta_2 = \|\mu_2\|^2$.} \quad (8.56)$$

Letting

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, I_{\nu_1+\nu_2} \right), \quad (8.57)$$

we have that

$$W_1 + W_2 = \|Z\|^2 \sim \chi^2_{\nu_1+\nu_2}(\|\mu_1\|^2 + \|\mu_2\|^2) = \chi^2_{\nu_1+\nu_2}(\Delta_1 + \Delta_2). \quad \Box \quad (8.58)$$

If both $W_i$’s are central chi-squares, then their sum is also central chi-squares. In fact, from (5.15), we have that a $\chi^2_{\nu}$ is $\Gammaamma(\frac{1}{2}, \frac{1}{2})$, and a $\chi^2_{\nu}$ is the sum of $\nu$ independent $\chi^2_{1}$’s, hence the sum of independent $\Gammaamma(\frac{1}{2}, \frac{1}{2})$’s from Example 5.2.2. Thus

$$\chi^2_{\nu} = \Gammaamma(\frac{\nu}{2}, \frac{1}{2}), \quad (8.59)$$

hence its pdf is

$$f_\nu(y) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{\nu/2-1} e^{-y/2}, \text{ $y > 0$.} \quad (8.60)$$
8.5.1 Moments of a chi-squared

The moments of a central chi-squared are easily obtained from those of a gamma. From (3.75) and (3.76), we have that the mean and variance of a $\Gamma(\alpha, \lambda)$ are $\alpha/\lambda$ and $\alpha/\lambda^2$, respectively, hence

$$E[\chi^2_\nu] = \nu \quad \text{and} \quad \text{Var}[\chi^2_\nu] = 2\nu. \quad (8.61)$$

Turn to the noncentral chi-squared. Because a $\chi^2_\nu(\Delta) = \chi^2_1(\Delta) + \chi^2_{\nu-1}$ where the latter two are independent,

$$E[\chi^2_\nu(\Delta)] = E[\chi^2_1(\Delta)] + E[\chi^2_{\nu-1}] \quad \text{and} \quad \text{Var}[\chi^2_\nu(\Delta)] = \text{Var}[\chi^2_1(\Delta)] + \text{Var}[\chi^2_{\nu-1}]. \quad (8.62)$$

Thus we need to find the moments of a $\chi^2_1(\Delta)$. Now $(\sqrt{\Delta} + X)^2 \sim \chi^2_1(\Delta)$, where $X \sim N(0, 1)$, hence

$$E[\chi^2_1(\Delta)] = E[(\sqrt{\Delta} + X)^2] = \Delta + 2\sqrt{\Delta}E[X] + E[X^2] = \Delta + 1. \quad (8.63)$$

Also,

$$E[(\chi^2_1(\Delta))^2] = E[(\sqrt{\Delta} + X)^4] = \Delta^4 + 4\Delta^{3/2}X + 6\Delta X^2 + 4\sqrt{\Delta}X^3 + X_4. \quad (8.64)$$

Now $E[X] = E[X^3] = 0, E[X^2] = 1$, and $E[X^4] = 3$ (can be found a number of ways), hence

$$E[(\chi^2_1(\Delta))^2] = \Delta^4 + 6\Delta + 3, \quad (8.65)$$

and

$$\text{Var}[(\chi^2_1(\Delta))^2] = \Delta^4 + 6\Delta + 3 - (\Delta + 1)^2 = 4\Delta + 2. \quad (8.66)$$

Putting (8.62) together with (8.64) and (8.65), we obtain

$$E[\chi^2_\nu(\Delta)] = \nu + \Delta \quad \text{and} \quad \text{Var}[\chi^2_\nu(\Delta)] = 2\nu + 4\Delta. \quad (8.67)$$

8.5.2 Other covariance matrices

What if $Z \sim N(\mu, \Sigma)$ for general $\Sigma$? We can still find a chi-squared distribution, but we first have to transform the $Z$ so that the covariance matrix is $I_\nu$. We will look at three possibilities, which are actually all special cases of using generalized inverses.

The simplest is if $Z \sim N(\mu, \sigma^2 I_\nu)$. Then

$$\frac{1}{\sigma} Z \sim N(\frac{1}{\sigma} \mu, I_\nu), \quad (8.68)$$

which means

$$\frac{\|Z\|^2}{\sigma^2} \sim \chi^2_\nu(\frac{\|\mu\|^2}{\sigma^2}), \quad \text{or} \quad \|Z\|^2 \sim \sigma^2 \chi^2_\nu(\frac{\|\mu\|^2}{\sigma^2}). \quad (8.69)$$

a “scaled” chi-square. Suppose more generally that $\Sigma$ is $\nu \times \nu$ and invertible. Then it has a symmetric square root, say $\Sigma^{1/2}$, which is also invertible. Multiply the vector by that inverse:

$$\Sigma^{-1/2} Z \sim N(\Sigma^{-1/2} \mu, \Sigma^{-1/2} \Sigma \Sigma^{-1/2}) = N(\Sigma^{-1/2} \mu, I_\nu). \quad (8.70)$$
Applying the definition yields
\[ \| \Sigma^{-1/2} Z \|^2 \sim \chi_v^2(\| \Sigma^{-1/2} \mu \|^2), \text{ or } Z' \Sigma^{-1} Z \sim \chi_v^2(\mu' \Sigma^{-1} \mu). \tag{8.71} \]

Also,
\[ (Z - \mu)' \Sigma^{-1} (Z - \mu) \sim \chi_v^2. \tag{8.72} \]

Finally, consider the \( H_n \) in (8.42). It is not positive definite, hence not invertible, because \( 1_n' H_n 1_n = 0 \). Thus we cannot immediately use (8.70). There is a nice result for symmetric idempotent \( n \times n \) covariance matrices \( H \). We apply it to \( HZ \), where \( Z \sim \mathcal{N}(\mu, I_n) \), so that
\[ HZ \sim \mathcal{N}(\mu, H). \tag{8.73} \]

We first look at the spectral decomposition, \( H = \Gamma \Lambda \Gamma' \), and note that
\[ H = \Gamma \Lambda \Gamma' = HH = \Gamma \Lambda \Gamma' \Gamma \Lambda' = \Gamma \Lambda^2 \Gamma', \tag{8.74} \]
so that \( \Lambda^2 = \Lambda \). Because \( \Lambda \) is diagonal, we have
\[ \lambda_i^2 = \lambda_i \text{ for all } i, \tag{8.75} \]
which means that each \( \lambda_i \) must be either 0 or 1. Because the trace of \( H \) is the sum of the eigenvalues, there must be \( \text{trace}(H) \) 1’s. These come first in the \( \Lambda \), hence
\[ \Lambda = \begin{pmatrix} I_v & 0 \\ 0 & 0 \end{pmatrix}, \quad v = \text{trace}(H). \tag{8.76} \]

In \( H \), the last \( n - v \) columns of \( \Gamma \) are multiplied by \( \lambda_i = 0 \), hence with \( \gamma_i \) being the \( i^{th} \) column of \( \Gamma \),
\[ H = \Gamma_1 \Gamma_1', \quad \Gamma_1 = (\gamma_1, \ldots, \gamma_v), n \times v. \tag{8.77} \]

Because the columns are orthonormal,
\[ \Gamma_1' \Gamma_1 = I_v. \tag{8.78} \]

Now look at
\[ \Gamma_1' HZ = \Gamma_1' Z \sim N(\Gamma_1' H \mu, \Gamma_1' H H \Gamma_1) = N(\Gamma_1' \mu, I_v) \tag{8.79} \]
because
\[ \Gamma_1' H = \Gamma_1' \Gamma_1 \Gamma_1' = \Gamma_1'. \tag{8.80} \]

The definition of chi-squares can be applied to (8.79):
\[ Z' \Gamma_1 \Gamma_1' Z \sim \chi_v^2(\mu' \Gamma_1 \Gamma_1' \mu), \tag{8.81} \]
i.e.,
\[ Z' HZ \sim \chi_v^2(\mu' H \mu), \quad v = \text{trace}(H). \tag{8.82} \]

Note that happily we do not need to explicitly find the eigenvectors \( \gamma_i \).

This result is used in liner models, where \( H \) is a projection matrix on a particular subspace, and \( Z' HZ \) are the sum-of-squares for that subspace.
Sum of squared deviations

Now back to Section 8.4, with

\[ X \sim N(\mu_1, \sigma^2 I_n), \quad U = \sum (X_i - X)^2 = X'H_nX. \] (8.83)

Here, \( H_n \mu_1 = 0_n \), and

\[ \frac{1}{\sigma} H_nX \sim N(0, H_n). \] (8.84)

Thus the \( U \) is central chi-squared:

\[ \frac{1}{\sigma^2} X'H_nX \sim \chi^2_v, \quad \nu = \text{trace}(H_n). \] (8.85)

The trace is not difficult:

\[ \text{trace}(H_n) = \text{trace}(I_n - \frac{1}{n} 1'n1_n) = \text{trace}(I_n) - \frac{1}{n} n = n - 1. \] (8.86)

We can now finish Section 8.4.

**Lemma 11.** If \( X_1, \ldots, X_n \) are iid \( N(\mu, \sigma^2) \), then \( X \) and \( \sum (X_i - \overline{X})^2 \) are independent,

\[ \overline{X} \sim N(\mu, \frac{1}{n} \sigma^2) \quad \text{and} \quad \sum (X_i - \overline{X})^2 \sim \sigma^2 \chi^2_{n-1}. \] (8.87)

### 8.6 Student’s \( t \) distribution

Here we answer the question of how to find a confidence interval for \( \mu \) as in (8.35) when \( \sigma \) is unknown. The "pivotal quantity" we use here is

\[ T = \frac{\overline{X} - \mu}{S/\sqrt{n}}, \quad S^2 \sim \frac{\sum (X_i - \overline{X})^2}{n - 1}. \] (8.88)

Its distribution is next.

**Definition 11.** Suppose \( Z \sim N(0, 1) \) and \( U \sim \chi^2_v \), where \( Z \) and \( U \) are independent. Then

\[ T = \frac{Z}{\sqrt{U/v}} \] (8.89)

is Student’s \( t \) on \( v \) degrees of freedom, written

\[ T \sim t_v. \] (8.90)

From Lemma 11, we have the conditions of Definition 11 satisfied by setting

\[ Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad U = \frac{\sum (X_i - \overline{X})^2}{\sigma^2}. \] (8.91)

Then

\[ T = \frac{(\overline{X} - \mu) / (\sigma/\sqrt{n})}{\sqrt{\frac{\sum (X_i - \overline{X})^2}{n - 1}}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}. \] (8.92)
A 95% confidence interval for \( \mu \) is then

\[
\bar{X} \pm t_{n-1,0.025} \frac{s}{\sqrt{n}},
\]

(8.93)

where \( t_{v,\alpha/2} \) is the cutoff point that satisfies

\[
P[-t_{v,\alpha/2} < t_v < t_{v,\alpha/2}] = 1 - \alpha.
\]

(8.94)

8.7 Linear models and the conditional distribution

Expanding on the simple linear model in (7.38), we consider the conditional model

\[
Y | X = x \sim N(\alpha + \beta x, \Sigma_e) \quad \text{and} \quad X \sim N(\mu_X, \Sigma_X),
\]

(8.95)

where \( Y \) is \( q \times 1 \), \( X \) is \( p \times 1 \), and \( \beta \) is a \( q \times p \) matrix. As in Example 7.7.1,

\[
E = Y - \alpha - \beta X \quad \text{and} \quad X
\]

are independent and multivariate normal, and their joint distribution is also multivariate normal:

\[
\left( \begin{array}{c} X \\ E \end{array} \right) \sim N \left( \begin{array}{c} \mu_X \\ \mu_Y \end{array} \right), \left( \begin{array}{cc} \Sigma_{XX} & 0 \\ 0 & \Sigma_e \end{array} \right) \). \quad (8.97)
\]

To find the joint distribution of \( X \) and \( Y \), we note that \((X',Y')'\) is a affine transformation of \((X',E')'\), hence is multivariate normal. Specifically,

\[
\left( \begin{array}{c} X \\ Y \end{array} \right) = \left( \begin{array}{c} 0_p \\ \alpha \end{array} \right) + \left( \begin{array}{cc} I_p & 0 \\ 0 & \beta \end{array} \right) \left( \begin{array}{c} X \\ E \end{array} \right) \\
\sim N \left( \begin{array}{c} \mu_X \\ \mu_Y \end{array} \right), \left( \begin{array}{cc} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY} & \Sigma_{YY} \end{array} \right),
\]

(8.98)

where

\[
\left( \begin{array}{c} \mu_X \\ \mu_Y \end{array} \right) = \left( \begin{array}{c} 0_p \\ \alpha \end{array} \right) + \left( \begin{array}{cc} I_p & 0 \\ 0 & \beta \end{array} \right) \left( \begin{array}{c} \mu_X \\ \mu_Y \end{array} \right) = \left( \begin{array}{c} \mu_X \\ \alpha + \beta \mu_X \end{array} \right)
\]

(8.99)

and

\[
\left( \begin{array}{cc} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YY} & \Sigma_{YY} \end{array} \right) = \left( \begin{array}{cc} I_p & 0 \\ \beta & I_q \end{array} \right) \left( \begin{array}{cc} \Sigma_{XX} & 0 \\ 0 & \Sigma_e \end{array} \right) \left( \begin{array}{cc} I_p & \beta' \\ \beta & I_q \end{array} \right) \\
= \left( \begin{array}{cc} \Sigma_{XX} & \Sigma_{XX} \beta' \\ \beta \Sigma_{XX} & \Sigma_e + \beta \Sigma_{XX} \beta' \end{array} \right).
\]

(8.100)

As we did for the bivariate normal, we invert the above process to find the conditional distribution of \( Y \) given \( X \) from the joint. First, we solve for \( \alpha \) and \( \beta \) in (8.100):

\[
\beta = \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XX} \quad \text{and} \quad \Sigma_e = \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}.
\]

(8.101)
Lemma 12. Suppose

\[
\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} \right),
\]

where \( \Sigma_{XX} \) is invertible. Then

\[
Y | X = x \sim N(\alpha + \beta x, \Sigma_e),
\]

where \( \alpha, \beta \) and \( \Sigma_e \) are given in (8.101).
Part II

Statistical Inference
Most, although not all, of the material so far has been straight probability calculations, that is, we are given a probability distribution, and try to figure out the implications (what $X$ is likely to be, marginals, conditionals, moments, etc.). Statistics generally concerns itself with the reverse problem, that is, observing the data $X = x$, and having then to guess aspects of the probability distribution that generated $x$. This “guessing” goes under the general rubric of inference. Four main aspects of inference are

- **Estimation**: What is the best guess of a particular parameter (vector), or function of the parameter? The estimate may be a point estimate, or a point estimate and measure of accuracy, or an interval or region, e.g., “The mean is in the interval (10.44, 19.77).”

- **Hypothesis testing**: The question is whether a specific hypothesis, the null hypothesis, about the distribution is true, so that the inference is basically either “yes” or “no”, along with an idea of how reliable the conclusion is.

- **Prediction**: One is interested in predicting a new observation, possibly depending on a covariate. For example, the data may consist of a number of $(X_i, Y_i)$ pairs, and a new observation comes along, where we know the $x$ but not the $y$, and wish to guess that $y$. We may be predicting a numerical variable, e.g., return on an investment, or a categorical variable, e.g., the species of a plant.

The boundaries between these notions are not firm. One can consider prediction to be estimation, or classification as an extension of hypothesis testing with more than two hypotheses. Whatever the goal, the first task is to specify the statistical model.

### 9.1 Statistical models

A probability model consists of a random $X$ in space $\mathcal{X}$ and a probability distribution $P$. A statistical model also has $X$ and space $\mathcal{X}$, but an entire family $\mathcal{P}$ of probability distributions on $\mathcal{X}$. By family we mean a set of distributions; the only restriction being that they are all distributions for the same $\mathcal{X}$. Such families can be quite general, e.g.,

$$\mathcal{X} = \mathbb{R}^n, \mathcal{P} = \{P \mid X_1, \ldots, X_n \text{ are iid with finite mean and variance}\}.$$  

(9.1)
This family includes all kinds of distributions (iid normal, gamma, beta, binomial), but not ones with the $X_i$’s correlated, or distributed Cauchy (which has no mean or variance). Another possibility is the family with the $X_i$’s iid with a continuous distribution.

Often, the families are parametrized by a finite-dimensional parameter $\theta$, i.e.,

$$\mathcal{P} = \{ P_\theta \mid \theta \in \Theta \}, \text{ where } \Theta \subset \mathbb{R}^K. \quad (9.2)$$

The $\Theta$ is called the **parameter space**. We are quite familiar with parameters, but for statistical models we must be careful to specify the parameter space as well. For example, suppose $X$ and $Y$ are independent, $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$. Then the following parameter spaces lead to distinctly different models:

$$\begin{align*}
\Theta_1 &= \left\{ (\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2) \in \mathbb{R} \times (0, \infty) \times \mathbb{R} \times (0, \infty) \right\} \\
\Theta_2 &= \left\{ (\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2) \mid \mu_X \in \mathbb{R}, \mu_Y \in \mathbb{R}, \sigma_X^2 = \sigma_Y^2 \in (0, \infty) \right\} \\
\Theta_3 &= \left\{ (\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2) \mid \mu_X \in \mathbb{R}, \mu_Y \in \mathbb{R}, \sigma_X^2 = \sigma_Y^2 = 1 \right\} \\
\Theta_4 &= \left\{ (\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2) \mid \mu_X \in \mathbb{R}, \mu_Y \in \mathbb{R}, \mu_X > \mu_Y, \sigma_X^2 = \sigma_Y^2 \in (0, \infty) \right\}. \quad (9.3)
\end{align*}$$

The first model places no restrictions on the parameters, other than the variances are positive. The second one demands the two variances be equal. The third sets the variances to 1, which is equivalent to saying that the variances are known to be 1. The last one equates the variances, as well as specifying that the mean of $X$ is larger than that of $Y$.

A **Bayesian model** includes a (prior) distribution on $\mathcal{P}$, which in the case of a parametrized model means a distribution on $\Theta$. In fact, the model could include a family of prior distributions, although we will not deal with that case explicitly.

Before we introduce inference, we take a brief look at how probability is interpreted.

### 9.2 Interpreting probability

In Section 2.1, we defined probability models mathematically, starting with some axioms. Everything else flowed from those axioms. But as in all mathematical objects, they do not in themselves have physical reality. In order to make practical use of the results, we must somehow connect the mathematical objects to the physical world. That is, how is one to interpret $P[A]$? In the case of games of chance, people generally feel confident that they know what “the chance of heads” or “the chance of a full house” mean. But other probabilities may be less obvious, e.g., “the chance that it rains next Tuesday” or “the chance ____ and ____ get married” (fill in the blanks with any two people). Two popular interpretations are **frequency** and **subjective**. Both have many versions, and there are also many other interpretations, but much of this material is beyond the scope of the author. Here are sketches of the two.

**Frequency.** An experiment is presumed to be repeatable, so that one could conceivably repeat the experiment under the exact same conditions over and over again (i.e., infinitely often). Then the probability of a particular event $A$, $P[A]$, is the long-run proportion of times it occurs, as the experiment is repeated forever. That is, it is the frequency $A$ occurs. This interpretation implies that probability is objective in the sense that it is inherent in the experiment, not a product of one’s beliefs. This interpretation works well for games of chance. One can imagine rolling a die or spinning
a roulette wheel an “infinite” number of times. Population sampling also fits in well, as one could imagine repeatedly taking a random sample of 100 subjects from a given population. The frequentist interpretation can not be applied to situations that are not in principle repeatable, such as whether two people will get married, or whether a particular candidate will win an election. One could imagine redoing the world over and over (like in the movie *Groundhogs Day*), but that is fiction.

Much of “classical” statistics is based on frequentist concepts. For example, a 95% confidence interval for the mean $\mu$ is a random interval with the property that, if the experiment were repeated over and over, about 95% of the intervals would contain $\mu$. That is, for $X \sim N(\mu, 1)$,

\[
P[X - 1.96 < \mu < X + 1.96] = 0.95 \text{ hence } (X - 1.96, X + 1.96) \text{ is a 95% confidence interval for } \mu.
\]  

(9.4)

Thus if we observe $X = 6.7$, say,

\[
(4.74, 8.66) \text{ is a 95% confidence interval for } \mu
\]

BUT $P[4.74 < \mu < 8.66] \neq 0.95$. (9.5)

The reason for the last statement is that there is nothing random in the event “4.74 < $\mu$ < 8.66”. The mean $\mu$ is the fixed but unknown population mean, and 4.74 and 8.66 are just numbers. Thus probability does not make sense there. Either $\mu$ is between 4.74 and 8.66, or it is not, we just do not know which. The randomness of $\mu$ does not come under the frequentist purview (unless of course a further repeatable experiment produces $\mu$, e.g., one first randomly chooses a population to sample from).

**Subjective.** The subjective approach allows each person to have a different probability, so that for a given person, $P[A]$ is that person’s opinion of the probability of $A$. The only assumption is that each person’s probabilities cohere, that is, satify the probability axioms. Subjective probability can be applied to any situation. One repeatable experiment, people’s subjective probabilities would tend to agree, whereas in other cases, such as the probability a certain team will win a particular game, their probabilities could differ widely. Subjectively, $P[4.74 < \mu < 8.66]$ does make sense; one drawback is that each person can have a different value.

Some subjectivists make the assumption that any given person’s subjective probabilities can be elicited using a betting paradigm. For example, suppose the event in question is “Pat and Leslie will get married,” the choices being “Yes” and “No”, and we wish to elicit your probability of the event “Yes”. We give you $10, and ask you for a number $w$, which will be used in two possible bets:

\[
\text{Bet 1 } \rightarrow \text{ Win } w \text{ if “Yes”, Lose } 10 \text{ if “No”,}
\]

\[
\text{Bet 2 } \rightarrow \text{ Lose } 10 \text{ if “Yes”, Win } \left(\frac{100}{w}\right) \text{ if “No”}
\]  

(9.6)

Some dastardly being will decide which of the bets you will take, so the $w$ should be an amount for which you are willing to take either of those two bets. For example, if you choose $w = 5$, then you are willing to accept a bet that pays only $5 if they do get married, and loses $10 if they don’t; AND you are willing to take a bet that wins $20 if they do not get married, and loses $10 if they do. These numbers suggest you expect they will get married. Suppose $p$ is your subjective probability of “Yes”. Then your willingness to take Bet 1 means you expect to not lose money:

\[
\text{Bet 1 } \rightarrow E[\text{Winnings}] = p \cdot w - (1 - p) \cdot 10 \geq 0.
\]  

(9.7)
Same with Bet 2:

\[ \text{Bet 2} \rightarrow E[\text{Winnings}] = -p \times ($10) + (1 - p) \times ($100/w) \geq 0. \]  
(9.8)

A little algebra translates those two inequations into

\[ p \geq \frac{10}{10 + w} \quad \text{and} \quad p \leq \frac{100/w}{10 + 100/w} = \frac{10}{10 + w}, \]  
(9.9)

which of course means that

\[ p = \frac{10}{10 + w}. \]  
(9.10)

With \( w = 5 \), your \( p = 2/3 \) that they will get married.

The betting approach is then an alternative to the frequency approach. Whether it is practical to elicit an entire probability distribution (i.e., \( P[A] \) for all \( A \subset X \)), and whether the result will satisfy the axioms, is questionable, but the main point is that there is in principle a grounding to a subjective probability.

9.3 Approaches to inference

Paralleling the interpretations of probability are the two main approaches to statistical inference: frequentist and Bayesian. Both aim to make inferences about \( \theta \) based on observing the data \( x = \bar{x} \), but take different tacks.

**Frequentist** The frequentist approach assumes that the parameter \( \theta \) is fixed but unknown (that is, we only know that \( \theta \in \Theta \)). An inference is an action, which is a function

\[ \delta : X \longrightarrow A, \]  
(9.11)

for some action space \( A \). The action space depends on the type of inference desired. For example, if one wishes to estimate \( \theta \), then \( \delta(x) \) would be the estimate, and \( A = \Theta \). Or \( \delta \) may be a vector containing the estimate as well as an estimate of its standard error, or it may be a two-dimensional vector representing a confidence interval. In hypothesis testing, we often take \( A = \{0, 1\} \), where 0 means accept the null hypothesis, and 1 means reject it. The properties of a procedure \( \delta \), which would describe how good it is, are based on the behavior if the experiment were repeated over and over, with \( \theta \) fixed. Thus an estimator \( \delta \) of \( \theta \) is **unbiased** if

\[ E_{\theta} [\delta(X)] = \theta \quad \text{for all} \quad \theta \in \Theta. \]  
(9.12)

Or a confidence interval procedure \( \delta(x) = (l(x), u(x)) \) has 95\% coverage if

\[ P_{\theta} [l(X) < \theta < u(X)] \geq 0.95 \quad \text{for all} \quad \theta \in \Theta. \]  
(9.13)

Understand that the 95\% does not refer to your particular interval, but rather to the infinite number of intervals that you imagine arising from repeating the experiment over and over.

**Bayesian.** The frequentist does not tell you what to think of \( \theta \). It just produces a number or numbers, then reassures you by telling you what would happen if you repeated the experiment an infinite number of times. The Bayesian approach, by contrast, tells you what to think. More precisely, given your prior distribution on
Θ, which may be your subjective distribution, the Bayes approach tells you how to update your opinion upon observing $X = x$. The update is of course the posterior, which we know how to find using Bayes Theorem 3. The posterior $f_{\theta|X}(\theta | x)$ is the inference, or at least all inferences are derived from it. For example, an estimate could be the posterior mean, median, or mode. A 95% probability interval is any interval $(l, u)$ such that

$$P[l < \theta < u \mid X = x] = 0.95.$$  
(9.14)

A hypothesis test would calculate

$$P[\text{Null hypothesis is true} \mid X = x],$$  
(9.15)

or, if an accept/reject decision is desired, reject the null hypothesis if the posterior probability of the null is less than 0.01, say.

A drawback to the frequentist approach is that we cannot say what we want, such as the probability a null hypothesis is true, or the probability $\mu$ is between two numbers. Bayesians can make such statements, but as the cost of having to come up with a (subjective) prior. The subjectivity means that different people can come to different conclusions from the same data. (Imagine a tobacco company and a consumer advocate analyzing the same smoking data.) Fortunately, there are more or less well-accepted "objective" priors, and especially when the data is strong, different reasonable priors will lead to practically the same posteriors. From an implementation point of view, sometimes frequentist procedures are computationally easier, and sometimes Bayesian procedures are. It may not be philosophically pleasing, but is not a bad idea to take an opportunistic view and use whichever approach best moves your understanding along.

There are other approaches to inference, such as the likelihood approach, the structural approach, the fiducial approach and the fuzzy approach. These are all interesting and valuable, but seem a bit iffy to me.

The rest of the course goes more deeply into inference.
10.1 Definition of estimator

We assume a model with parameter space $\Theta$, and suppose we wish to estimate some function $g$ of $\theta$,  
\[ g : \Theta \rightarrow \mathbb{R}. \tag{10.1} \]
This function could be $\theta$ itself, or one component of $\theta$. For example, if $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$, some possible $g$’s are
\[
\begin{align*}
g(\mu, \sigma^2) &= \mu; \\
g(\mu, \sigma^2) &= \sigma; \\
g(\mu, \sigma^2) &= \sigma/\mu = \text{coefficient of variation}; \\
g(\mu, \sigma^2) &= P[X_i \leq 10] = \Phi((10 - \mu)/\sigma), \tag{10.2}
\end{align*}
\]
where $\Phi$ is the distribution function for $N(0, 1)$.

An estimator is a function $\delta(X)$, 
\[ \delta : \mathcal{X} \rightarrow A, \tag{10.3} \]
where $A$ is some space, presumably the space of $g(\theta)$, but not always. The hope is that $\delta(X)$ is close to $g(\theta)$. The estimator can be any function of $X$, but cannot depend on an unknown parameter. Thus with $g(\mu, \sigma^2) = \sigma/\mu$ in the above example,
\[
\begin{align*}
\delta(x_1, \ldots, x_n) &= \frac{s}{\bar{x}} \quad [s^2 = \frac{\sum(x_i - \bar{x})^2}{n}] & \text{is an estimator,} \\
\delta(x_1, \ldots, x_n) &= \frac{\sigma}{\bar{x}} & \text{is not an estimator.} \tag{10.4}
\end{align*}
\]

Any function can be an estimator, but that does not mean it will be a particularly good estimator. There are basically two questions we must address: How does one find reasonable estimators? How do we decide which estimators are good? As may be expected, the answers differ depending on whether one is taking a frequentist approach or a Bayesian approach.

This chapter will introduce several methods of estimation. Section 12.3 in the next chapter delves into a certain optimality for estimators.
10.2 Plug-in methods

Suppose the data consist of $X_1, \ldots, X_n$ iid vectors, each with space $\mathcal{X}_0$ and distribution function $F$. Many parameters, such as the mean or variance, are defined for wide classes of distributions $F$. The simplest notion of a plug-in estimator is to use the sample version of the population parameter, so that $\bar{X}$ is an estimator for $\mu$, and $S^2$, the sample variance, is an estimator for $\sigma^2$, and an estimator of $P[X_i \leq 10]$ is

$$\delta(x) = \frac{\# \{ x_i \leq 10 \}}{n}. \quad (10.5)$$

Such estimators can be formalized by considering a large family $\mathcal{F}$ of distribution functions, which includes at least all the distributions with finite $\mathcal{X}_0$, and functionals on $\mathcal{F}$,

$$\theta : \mathcal{F} \rightarrow \mathbb{R}. \quad (10.6)$$

That is, the argument for the function is $F$. The mean, median, variance, $P[X_i \leq 10] = F(10)$, etc., are all such functionals for appropriate $F$’s. Notice that the $\alpha$ and $\beta$ of a beta distribution are not such functionals, because they are not defined for non-beta distributions. The $\theta(F)$ is then the population parameter $\theta$. The sample version is

$$\hat{\theta} = \theta(\hat{F}_n), \quad (10.7)$$

where $\hat{F}_n$ is the empirical distribution function for a given sample $x_1, \ldots, x_n$, defined by

$$\hat{F}_n(x) = \frac{\# \{ x_i \mid x_i \leq x \}}{n} \quad (10.8)$$

in the univariate case, or

$$\hat{F}_n(x) = \frac{\# \{ x_i \mid x_{ij} \leq x_j \text{ for all } j = 1, \ldots, p \}}{n} \quad (10.9)$$

if the data are $p$-variate:

$$\bar{x}_i = \left( \begin{array}{c} x_{i1} \\ \vdots \\ x_{ip} \end{array} \right) \quad \text{and} \quad \bar{x} = \left( \begin{array}{c} x_1 \\ \vdots \\ x_p \end{array} \right) \in \mathbb{R}^p. \quad (10.10)$$

This $\hat{F}_n$ is the distribution function for the random vector $\bar{X}^*$ which has space

$$\mathcal{X}_0^* = \{ \text{the distinct values among } \bar{x}_1, \ldots, \bar{x}_n \} \quad (10.11)$$

and probabilities

$$P^* [\bar{X}^* = \bar{x}^*] = \frac{\# \{ \bar{x}_i \mid \bar{x}_i = \bar{x}^* \}}{n}, \quad \bar{x}^* \in \mathcal{X}_0^*. \quad (10.12)$$

Thus $\bar{X}^*$ is generated by randomly choosing one of the observations $\bar{x}_i$. For example, if the sample is $3, 5, 2, 6, 2, 2$, then

$$\mathcal{X}_0^* = \{2, 3, 5, 6\}, \quad \text{and} \quad P[\bar{X}^* = 2] = \frac{1}{2}, P[\bar{X}^* = 3] = P[\bar{X}^* = 5] = P[\bar{X}^* = 6] = \frac{1}{6}. \quad (10.13)$$
The $F_n$ is an estimate of $F$, so the estimate (10.7) for $\theta$ is found by plugging in $F_n$ for $F$. If $\theta$ is the mean, then $\theta(F_n) = \bar{x}$, and if $\theta$ is the variance, $\theta(F_n) = s^2 = \sum(x_i - \bar{x})^2 / n$. As in (10.2), the coefficient of variation can be estimated by the sample version, $s/\bar{x}$.

Other plug-in estimators arise in parametric models, where one has estimates of the parameters and wishes to estimate some function of them, or one has estimates of some functions of the parameters and wishes to estimate the parameters themselves. For example, if the $X_i$'s are iid $N(\mu, \sigma^2)$, $P[X_i \leq 10] = \Phi((10 - \mu)/\sigma)$, where $\Phi$ is the distribution function of the standard normal. Then instead of (10.5), we could use the estimate

$$P[X_i \leq 10] = \Phi\left(\frac{10 - \bar{x}}{s}\right). \tag{10.14}$$

Or suppose the $X_i$'s are iid Beta$(\alpha, \beta)$, with $(\alpha, \beta) \in (0, \infty) \times (0, \infty)$. Then from (6.30) and (6.31), the population mean and variance are

$$\mu = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{10.15}$$

Then $\bar{x}$ and $s^2$ are estimates of those functions of $\alpha$ and $\beta$, hence the estimates $\hat{\alpha}$ and $\hat{\beta}$ of $\alpha$ and $\beta$ would be the solutions to

$$\bar{x} = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} \quad \text{and} \quad s^2 = \frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha} + \hat{\beta})^2(\hat{\alpha} + \hat{\beta} + 1)}, \tag{10.16}$$

or after some algebra,

$$\hat{\alpha} = \bar{x} \left(\frac{\bar{x}(1 - \bar{x})}{s^2} - 1\right) \quad \text{and} \quad \hat{\beta} = (1 - \bar{x}) \left(\frac{\bar{x}(1 - \bar{x})}{s^2} - 1\right). \tag{10.17}$$

The estimators in (10.17) are special plug-in estimators, called method of moments estimators, because the estimates of the parameters are chosen to match the population moments with their sample versions. The method of moments are not necessarily strictly defined. For example, in the Poisson$(\lambda)$, the mean and variance are $\lambda$, so that $\lambda$ could be $\bar{x}$ or $s^2$. Also, one has to choose moments that work. For example, if the data are iid $N(0, \sigma^2)$, and we wish to estimate $\sigma$, the mean is useless because one cannot do anything to match $0 = \bar{x}$.

### 10.3 Least squares

Another general approach to estimation chooses the estimate of $\theta$ so that the observed data are close to their expected values. For example, if one is trying to estimate the mean $\mu$ of iid $X_i$, the estimate $\hat{\mu}$ is chosen so that it is "close" to the $X_i$'s. It cannot be exactly equal to all the observations (unless they are equal), so some overall measure of closeness is needed. The most popular is least squares, that is, $\hat{\mu}$ is the value $a$ that minimizes

$$L(a; x_1, \ldots, x_n) = \sum(x_i - a)^2 \quad \text{over} \quad a. \tag{10.18}$$

This $L$ is easily minimized by differentiating with respect to $a$ then setting to 0, to obtain $a = \bar{x}$, that is, the sample mean is the least squares estimate of $\mu$. Measures other than squared error are useful as well. For example, least absolute deviations
minimizes $\sum |x_i - a|$, the minimizer being the sample median; a measure from support vector machine methods allows a buffer zone:

$$L(a; x_1, \ldots, x_n) = \sum (x_i - a)^2 I_{|x_i - a| < \epsilon},$$

(10.19)

that is, it counts only those observations less than $\epsilon$ away from $a$.

Least square is especially popular in linear models, where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent,

$$E[Y_i \mid X_i = x_i] = a + \beta x_i, \quad \text{Var}[Y_i] = \sigma^2_e.$$  

(10.20)

Then the least squares estimates of $\alpha$ and $\beta$ are the $a$ and $b$, respectively, that minimize

$$L(a, b; (x_1, y_1), \ldots, (x_n, y_n)) = \sum (y_i - a - bx_i)^2.$$  

(10.21)

Differentiating yields the normal equations,

$$\left( \begin{array}{c} \sum y_i \\ \sum x_i y_i \end{array} \right) = \left( \begin{array}{c} n \\ \sum x_i \end{array} \right) \left( \begin{array}{cc} a \\ b \end{array} \right),$$

(10.22)

which have a unique solution if and only if $\sum(x_i - \bar{x})^2 > 0$. (Which makes sense, because if all the $x_i$'s are equal, it is impossible to estimate the slope of the line through the points $(x_i, y_i)$.) Inverting the matrix, we have

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad \text{and} \quad a = \bar{y} - b \bar{x}.$$  

(10.23)

Note that these estimates are sample analogs of the $\alpha$ and $\beta$ in (8.101).

### 10.4 The likelihood function

If we know $\theta$, then density tells us what $X$ is likely to be. In statistics, we do not know $\theta$, but we do observe $X = x$, and wish to know what values $\theta$ is likely to be. The analog to the density for the statistical problem is the likelihood function, which is the same as the density, but considered as a function of $\theta$ for fixed $x$. The function is not itself a density (usually), because there is no particular reason to believe that the integral over $\theta$ for fixed $x$ is 1. Rather, the likelihood function gives the relative likelihood of various values of $\theta$.

For example, suppose $X \sim Binomial(n, \theta), \theta \in (0, 1), n = 5$. The pmf is

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \quad x = 0, \ldots, n.$$  

(10.24)

The likelihood is

$$L(\theta \mid x) = c_x \theta^x (1 - \theta)^{n-x}, \quad 0 < \theta < 1.$$  

(10.25)

Here, $c_x$ is a constant that may depend on $x$ but not on $\theta$. Suppose the observed value of $X$ is $x = 2$. Then one can find the relative likelihood of different values of $\theta$ from $L$, that is,

"relative likelihood of $\theta = .2$ to $\theta = .4" = \frac{L(.4; 2)}{L(.2; 2)} = \frac{(.4)^2(.6)^3}{(.2)^2(.8)^3} = \frac{27}{16} = 1.6875.$$  

(10.26)
10.4. The likelihood function

That is, based on seeing \( x = 2 \) heads in \( n = 5 \) flips of a coin, it is about 1.7 times as likely that the true probability of heads is \( \theta = .4 \) versus \( \theta = .2 \). Notice that the “\( c_x \)” in the likelihood (10.25) is irrelevant, because we look at ratios.

Likelihood is not probability, in particular because \( \theta \) is not necessarily random. Even if \( \theta \) is random, the likelihood is not the pdf, but rather the part of the pdf that depends on \( x \). That is, suppose \( \pi \) is the prior pdf of \( \theta \). Then the posterior pdf is

\[
f(\theta | x) = \frac{f(x | \theta)\pi(\theta)}{\int_{\Theta} f(x | \theta^*)\pi(\theta^*)d\theta^*} = c_x L(\theta; x)\pi(\theta). \quad (10.27)
\]

(Note: These densities should have subscripts, e.g., \( f(x | \theta) \) should be \( f_{X|\theta}(x | \theta) \), but I hope leaving them off is not too confusing.) Here, the \( c_x \) takes into account the \( c_x \) in (10.25) as well as the integral in the denominator of (10.27) (which does not depend on \( \theta \) — it is integrated away). Thus though the likelihood is not a density, it does tell us how to update the prior to obtain the posterior.

10.4.1 Maximum likelihood estimation

If \( L(\theta; x) \) reveals how likely \( \theta \) is in light of the data \( x \), it seems reasonable that the most likely \( \theta \) would be a decent estimate of \( \theta \). In fact, it is reasonable, and the resulting estimator is quite popular.

**Definition 12.** Given the model with likelihood \( L(\theta; x) \) for \( \theta \in \Theta \), the **maximum likelihood estimate (MLE)** at observation \( x \) is the unique value of \( \theta \) that maximizes \( L(\theta; x) \) over \( \theta \in \Theta \), if such unique value exists. Otherwise, the MLE does not exist at \( x \).

There are times when the likelihood does not technically have a maximum, but there is an obvious limit of the \( \theta \)'s that approach the supremum. For example, sup-
pose $X \sim \text{Uniform}(0, \theta)$. Then the likelihood at $x > 0$ is

$$L(\theta; x) = \frac{1}{\theta} I_{\{x < \theta\}}(\theta) = \begin{cases} \frac{1}{\theta} & \text{if } \theta > x \\ 0 & \text{if } \theta \leq x \end{cases}.$$  \hspace{1cm} (10.28)

For example, here is a plot of the likelihood when $x = 4$.

\begin{center}
\textbf{Uniform likelihood}
\end{center}

The highest point occurs at $\theta = 4$, almost. To be precise, there is no maximum, because the graph is not continuous at $\theta = 4$. But we still take the MLE to be $\hat{\theta} = 4$ in this case, because we could switch the filled-in dot from $(4, 0)$ to $(4, 1/4)$ by taking $X \sim \text{Uniform}(0, \theta]$, so that the pdf at $x = \theta$ is $1/\theta$, not 0.

Often, the MLE is found by differentiating the likelihood, or the log of the likelihood (called the loglikelihood). Because the log function is strictly increasing, the same value of $\theta$ maximizes the likelihood and the loglikelihood. For example, in the binomial example, from (10.25) the loglikelihood is (dropping the $c_x$)

$$l(\theta; x) = \log(L(\theta; x)) = x \log \theta + (n - x) \log(1 - \theta).$$  \hspace{1cm} (10.29)

Then

$$l'(\theta; x) = \frac{x}{\theta} - \frac{n - x}{1 - \theta}.$$  \hspace{1cm} (10.30)

Set that expression to 0 and solve for $\theta$ to obtain

$$\hat{\theta} = \frac{x}{n}.$$  \hspace{1cm} (10.31)

From Figure 10.1, one can see that the maximum is at $x/n = 0.4$.

If $\theta$ is $p \times 1$, then one must do a $p$-dimensional maximization. For example, suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. Then the loglikelihood is
10.5 The posterior distribution

(dropping the $\sqrt{2\pi}$'s),

$$\log(L(\mu, \sigma^2; x_1, \ldots, x_n)) = -\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 - \frac{n}{2} \log(\sigma^2). \quad (10.32)$$

We could go ahead and differentiate with respect to $\mu$ and $\sigma^2$, obtaining two equations. But notice that $\mu$ appears only in the sum of squares part, and we have already, in (10.18), found that $\hat{\mu} = \bar{x}$. Then for the variance, we need to maximize

$$\log(L(\bar{x}, \sigma^2; x_1, \ldots, x_n)) = -\frac{1}{2\sigma^2} \sum (x_i - \bar{x})^2 - \frac{n}{2} \log(\sigma^2). \quad (10.33)$$

Differentiating with respect to $\sigma^2$,

$$\frac{\partial}{\partial (\sigma^2)} \log(L(\bar{x}, \sigma^2; x_1, \ldots, x_n)) = \frac{1}{2\sigma^4} \sum (x_i - \bar{x})^2 - \frac{n}{2} \frac{1}{\sigma^2}, \quad (10.34)$$

and setting to 0 leads to

$$\hat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2}{n} = s^2. \quad (10.35)$$

So the MLE of $(\mu, \sigma^2)$ is $(\bar{x}, s^2)$.

### 10.5 The posterior distribution

In the Bayesian framework, the inference is the posterior, because the posterior has all the information about the parameter given the data. Estimation of $\theta$ then becomes a matter of summarizing the distribution. In Example 7.6.2, if $X \mid \theta \sim \text{Binomial}(n, \theta)$, and the prior is $\theta \sim \text{Beta}(\alpha, \beta)$, then

$$\theta \mid X = x \sim \text{Beta}(\alpha + x, \beta + n - x). \quad (10.36)$$

One estimate is the posterior mean,

$$E[\theta \mid X = x] = \frac{\alpha + x}{\alpha + \beta + n}. \quad (10.37)$$

Another is the posterior mode, the value with the highest density, which turns out to be

$$\text{Mode}[\theta \mid X = x] = \frac{\alpha + x - 1}{\alpha + \beta + n - 2}. \quad (10.38)$$

Note that both of these estimates are modifications of the MLE, $x/n$, and are close to the MLE if the parameters $\alpha$ and $\beta$ are small. One can think of $\alpha$ as the prior number of successes and $\alpha + \beta$ as the prior sample size; the larger $\alpha + \beta$, the more weight the prior has in the posterior.

A 95% probability interval for the parameter could be $(l, u)$ where

$$P[\theta < l \mid X = x] = P[\theta > u \mid X = x] = 0.025, \quad (10.39)$$

or commonly the $l$ and $u$ are chosen so that the interval is shortest, but still with 95% probability.
10.6 Bias, variance, and MSE

Rather than telling one what to think about the parameter, frequentist inference makes a statement, e.g., an estimate, confidence interval, or acceptance/rejection of a null hypothesis. The properties of the statement do not refer to the particular parameter at hand, but rather how the statement would behave if the experiment were repeated over and over. Thus the “95%” in the 95% confidence interval means that if the same procedure were repeated over and over again in similar experiments, about 95% of the resulting intervals would contain the parameter. This 95% coverage rate is an operating characteristic.

In estimation, any function can be an estimator, but what is a good one? There are many operating characteristics, but here we will present just three: bias, variance, and mean square error.

**Definition 13.** The bias of an estimator $\delta$ of $g(\theta)$ is the function

$$\text{Bias}(\theta; \delta) = E_\theta[\delta(X)] - g(\theta), \; \theta \in \Theta.$$  \hfill (10.40)

The estimator is unbiased if

$$E_\theta[\delta(X)] = g(\theta), \; \text{for all} \; \theta \in \Theta.$$  \hfill (10.41)

An unbiased estimator is then one that, on average, will equal the quantity being estimated (the estimand). Generally, bias is not good, but whether we have to worry about absolutely 0 bias, or just a little, is questionable. A good estimator will be close the estimand, and being unbiased does not necessarily mean being close, e.g., half the time the estimator could be extremely low, and half extremely high, but the lows and highs balance out. Thus variance is also a key operating characteristic: A low variance means the estimator is likely to be near its expected value, which is the estimand if the estimator is unbiased. Alternatively, one could look at closeness directly, e.g., in a squared-error sense.

**Definition 14.** The mean squared error (MSE) of an estimator $\delta$ of $g(\theta)$ is

$$\text{MSE}(\theta; \delta) = E_\theta[(\delta(X) - g(\theta))^2].$$  \hfill (10.42)

The MSE combines bias and variance in a handy way. Letting $e_\delta(\theta) = E_\theta[\delta(X)]$,

$$\text{MSE}(\theta; \delta) = E_\theta[(\delta(X) - g(\theta))^2]$$

$$= E_\theta[(\delta(X) - e_\delta(\theta) - (g(\theta) - e_\delta(\theta)))^2]$$

$$= E[(\delta(X) - e_\delta(\theta))^2] + 2E[(\delta(X) - e_\delta(\theta))(g(\theta) - e_\delta(\theta))]$$

$$+ E[(g(\theta) - e_\delta(\theta))^2]$$

$$= \text{Var}_\theta[\delta(X)] + \text{Bias}(\theta; \delta)^2,$$  \hfill (10.43)

because $g(\theta)$ and $e_\delta(\theta)$ are constants as far as $X$ is concerned. Thus MSE balances bias and variances.

For an example, consider again $X \sim \text{Binomial}(n, \theta)$, where we wish to estimate $\theta \in (0, 1)$. We have the MLE, $\hat{\delta}(x) = x/n$, and the Bayes posterior means, (10.37), for various $\alpha$ and $\beta$. To assess their operating characteristics, we find for the MLE,

$$E_{\theta} \left[ \frac{X}{n} \right] = \theta \; \text{and} \; \text{Var}_{\theta} \left[ \frac{X}{n} \right] = \frac{\theta(1 - \theta)}{n},$$  \hfill (10.44)
10.6. Bias, variance, and MSE

and for prior $\theta \sim Beta(\alpha, \beta)$,

$$E_\theta \left[ \frac{\alpha + X}{\alpha + \beta + n} \right] = \frac{\alpha + n\theta}{\alpha + \beta + n}$$

and

$$Var_\theta \left[ \frac{\alpha + X}{\alpha + \beta + n} \right] = \frac{n\theta(1-\theta)}{(n + \alpha + \beta)^2}. \quad (10.45)$$

(Note that formally, the MLE is the posterior mean if $\alpha = \beta = 0$. But that prior is not legitimate, that is, the parameters have to be positive in the beta.) We can see that the MLE is unbiased\(^1\), but the posterior means are biased, no matter what the parameters. On the other hand, the variance of the posterior means are always less than that of the MLE. Figure 10.2 graphs the MSE’s for the MLE, and the posterior means for $(\alpha, \beta) = (1, 1), (10, 10),$ and $(10, 3).$ Note that they are quite different, and that none of them is always best, i.e., lowest. The Bayes estimates tend to be good near the prior means, $1/2, 1/2,$ and $10/13,$ respectively. The Bayes estimate with $(\alpha, \beta) = (1, 1)$ is closest to having a flat, and fairly low, MSE. It is never very bad, though rarely the best.

---

\(^1\)It is by no means a general rule that MLE’s are unbiased.
Which estimator to use depends on where you wish the best MSE. (Unless you insist on unbiasedness, in which case you are stuck with \( \frac{x}{n} \).) If you use a beta prior to specify where you think the \( \theta \) is, then the corresponding posterior mean is the best, as we shall see in the next section.

Chapter 12 goes into unbiasedness more deeply. Chapter 14 presents the formal decision-theoretic approach.

### 10.7 Functions of estimators

If \( \hat{\theta} \) is an estimator of \( \theta \), then \( g(\hat{\theta}) \) is an estimator of \( g(\theta) \). Do the properties of \( \hat{\theta} \) transfer to \( g(\hat{\theta}) \)? Maybe, maybe not. Although there are exceptions (such as when \( g \) is linear for the first two statements), generally

- If \( \hat{\theta}_U \) is an unbiased estimator of \( \theta \) then \( g(\hat{\theta}_U) \) is not an unbiased estimator of \( g(\theta) \);
- If \( \hat{\theta}_B \) is the Bayes posterior mean of \( \theta \) then \( g(\hat{\theta}_B) \) is not the Bayes posterior mean of \( g(\theta) \);
- If \( \hat{\theta}_{ml} \) is the MLE of \( \theta \) then \( g(\hat{\theta}_{ml}) \) is the MLE of \( g(\theta) \).

The basic reason for the first two statements is the following:

**Theorem 6.** If \( Y \) is a random variable with finite mean, and \( g \) is a function of \( y \), then

\[
E[g(Y)] \neq g(E[Y])
\]

unless

- \( g \) is linear: \( g(y) = a + b \, y \) for constants \( a \) and \( b \);
- \( Y \) is essentially constant: \( P[Y = \mu] = 1 \) for some \( \mu \);
- You are lucky.

A simple example is \( g(x) = x^2 \). If \( E[X^2] = (E[X])^2 \), then \( X \) has variance 0. As long as \( X \) is not a constant, \( E[X^2] > (E[X])^2 \).

#### 10.7.1 Example: Poisson

Suppose \( X \sim \text{Poisson}(\theta) \), \( \theta \in (0, \infty) \). Then for estimating \( \theta \), \( \hat{\theta}_U = X \) is unbiased, and happens to be the MLE as well. For a Bayes estimate, take the prior \( \theta \sim \text{Exponential}(1) \). The likelihood is \( L(\theta; x) = e^{-\theta} \theta^x \), hence the posterior is

\[
\pi(\theta \mid x) = c.L(\theta; x)\pi(\theta) = c.e^{-\theta} \theta^x e^{-\theta} = c.\theta^x e^{-2\theta},
\]

which is \( \text{Gamma}(\alpha = x + 1, \lambda = 2) \). Thus from (3.75), the posterior mean with respect to this prior is

\[
\hat{\theta}_B = E[\theta \mid X = x] = \frac{x + 1}{2}.
\]
Now consider estimating \( g(\theta) = e^{-\theta} \), which is the \( P_{\theta}[X = 0] \). (So if \( \theta \) is the average number of telephone calls coming in an hour, \( e^{-\theta} \) is the chance there are 0 calls in the next hour.) Is \( g(\hat{\theta}_U) \) unbiased? No:

\[
E[g(\hat{\theta}_U)] = E[e^{-X}] = e^{-\theta} \sum_{x=0}^{\infty} \frac{(e^{-1}\theta)^x}{x!} = e^{-\theta} e^{-e^{-1}\theta} = e^{\theta(e^{-1}-1)} \neq e^{-\theta}.
\]

There is an unbiased estimator for \( g \), namely \( I[X = 0] \).

Turn to the posterior mean of \( g(\theta) \), which is \( \hat{g}(\theta_B) = E[g(\theta) \mid X = x] \)

\[
\hat{g}(\theta_B) = \int_0^\infty g(\theta) \pi(\theta \mid x) d\theta = \frac{2x+1}{\Gamma(x+1)} \int_0^\infty e^{-\theta} \theta^x e^{-2\theta} d\theta = \frac{2x+1}{\Gamma(x+1)} \int_0^\infty \theta^x e^{-3\theta} d\theta = \frac{2x+1}{
\frac{\Gamma(x+1)}{3^{x+1}}} = \left( \frac{2}{3} \right)^{x+1} \neq e^{-\hat{\theta}_B} = e^{-\frac{x+1}{2}}.
\]

For the MLE, we have to reparametrize. That is, with \( \tau = g(\theta) = e^{-\theta} \), the likelihood and loglikelihood are

\[
L^*(\tau; x) = \tau(-\log(\tau))^x \quad \text{and} \quad \log(L^*(\tau; x)) = \log(\tau) + x \log(-\log(\tau)).
\]

Then

\[
\frac{\partial}{\partial \tau} \log(L^*(\tau; x)) = \frac{1}{\tau} + \frac{x}{\tau \log(\tau)} = 0 \Rightarrow \hat{\tau}_{mle} = e^{-x} = e^{-\hat{\theta}_{mle}}.
\]

The MLE result follows in general for one-to-one functions because the likelihoods are related via

\[
L(\theta; x) = L^*(g(\theta); x),
\]

hence

\[
L(\hat{\theta}; x) > L(\theta; x) \quad \Rightarrow \quad L^*(g(\hat{\theta}); x) > L^*(g(\theta); x),
\]

so if \( \hat{\theta} \) maximizes the \( L \), \( g(\hat{\theta}) \) maximizes the \( L^* \).
Chapter 11

Likelihood and sufficiency

11.1 The likelihood principle

The likelihood function is introduced in Section 10.4. It is fundamental to Bayesian inference, and extremely useful in frequentist inference. It encapsulates the relationship between $\theta$ and the data. In this section we look at likelihood more formally.

Definition 15. Suppose $X$ has density $f(x \mid \theta)$ for $\theta \in \Theta$. Then a likelihood function for observation $X = x$ is

$$L(\theta; x) = c_x f(x \mid \theta), \quad \theta \in \Theta,$$  \hspace{1cm} (11.1)

where $c_x$ is any positive constant.

Two likelihood functions are the same if they are proportional.

Definition 16. Suppose the two models, $X$ with density $f(x \mid \theta)$ and $Y$ with density $g(y \mid \theta)$, depend on the same parameter $\theta$ with space $\Theta$. Then $x$ and $y$ have the same likelihood if for some positive constants $c_x$ and $c_y$ in (11.1),

$$L(\theta; x) = L^*(\theta; y) \text{ for all } \theta \in \Theta,$$  \hspace{1cm} (11.2)

where $L$ and $L^*$ are their respective likelihoods.

The two models in Definition 16 could very well be the same, in which case $x$ and $y$ are two possible elements of $X$. As an example, suppose the model has $X = (X_1, X_2)$, where the elements are iid $N(\mu, 1)$, and $\mu \in \Theta = \mathbb{R}$. Then the pdf is

$$f(x \mid \mu) = \frac{1}{2\pi} e^{-\frac{1}{2} \sum(x_i - \mu)^2} = \frac{1}{2\pi} e^{-\frac{1}{2} (x_1^2 + x_2^2) + (x_1 + x_2)\mu - \mu^2}. \hspace{1cm} (11.3)$$

Consider two possible observed vectors, $\underline{x} = (1, 2)$ and $\underline{y} = (-3, 6)$. These observations have quite different pdf values. Their ratio is

$$\frac{f(1, 2 \mid \mu)}{f(-3, 6 \mid \mu)} = 485165195. \hspace{1cm} (11.4)$$
This ratio is interesting in two ways: It shows the probability of being near \( x \) is almost half a billion times larger than the probability of being near \( y \), and it shows the ratio does not depend on \( \mu \). That is, their likelihoods are the same:

\[
L(\mu; x) = c_{(1,3)} e^{3\mu - \mu^2} \quad \text{and} \quad L(\mu; y) = c_{(-3,6)} e^{3\mu - \mu^2}.
\]

(11.5)

It does not matter what the constants are. We could just take \( c_{(1,3)} = c_{(-3,6)} = 1 \), but the important aspect is that the sums of the two observations are both 3, and the sum is the only part of the data that hits \( \mu \).

### 11.1.1 Binomial and negative binomial

The models can be different, but they do need to share the same parameter. For example, suppose we have a coin with probability of heads being \( \theta \in (0, 1) \), and we intend to flip it a number of times independently. Here are two possible experiments:

- **Binomial.** Flip the coin \( n = 10 \) times, and count \( X \), the number of heads, so that \( X \sim \text{Binomial}(10, \theta) \).

- **Negative binomial.** Flip the coin until there are 4 heads, and count \( Y \), the number of tails obtained. This \( Y \) is called \( \text{Negative Binomial}(4, \theta) \).

The space of the \( \text{Negative Binomial}(k, \theta) \) is \( Y = \{0, 1, 2, \ldots\} \), and the pmf can be found by noting that to obtain \( Y = y \), there must be exactly \( k - 1 \) heads among the first \( k - 1 + y \) flips, then the \( k^{th} \) head on the \( (k + y)^{th} \) flip:

\[
g(y \mid \theta) = \left[ \binom{k - 1 + y}{k - 1} \theta^{k-1}(1-\theta)^y \right] \times \theta = \binom{k - 1 + y}{k - 1} \theta^k(1-\theta)^y. \tag{11.6}
\]

Next, suppose we perform the binomial experiment and obtain \( X = 4 \) heads out of \( 10 \) flips. The likelihood is the usual binomial one:

\[
L(\theta ; 4) = \binom{10}{4} \theta^4(1-\theta)^6. \tag{11.7}
\]

Also, suppose we perform the negative binomial experiment and happen to see \( Y = 6 \) tails before the \( k = 4^{th} \) head. The likelihood here is

\[
L^*(\theta ; 6) = \binom{9}{3} \theta^4(1-\theta)^6. \tag{11.8}
\]

The likelihoods are the same. I left the constants there to illustrate that the pmf’s are definitely different, but erasing the constants leaves the same \( \theta^4(1-\theta)^6 \). These two likelihoods are based on different random variables. The binomial has a fixed number of flips but could have any number of heads (between 0 and 10), while the negative binomial has a fixed number of heads but could have any number of flips (over 4). In particular, either experiment with the given outcome would yield the same posterior for \( \theta \), and the same MLE for \( \theta \).

The **likelihood principle** says that if two outcomes (whether they are from the same experiment or not) have the same likelihood, then any inference made about \( \theta \) based on the outcomes must be the same. Any inference that is not the same under
the two scenarios are said to violate the likelihood principle. Bayesian inference does not violate the likelihood principle, nor does maximum likelihood estimation, as long as your inference is just “Here is the estimate ...”

Unbiased estimation does violate the likelihood principle. Keeping with the above example, we know that \( X/n \) is an unbiased estimator of \( \theta \) for the binomial. For \( Y \sim \text{Negative Binomial}(k, \theta) \), the unbiased estimator is found by ignoring the last flip, because that we know is always heads, so would bias the estimate if used. That is,

\[
E[\hat{\theta}_U] = \theta, \quad \hat{\theta}_U = \frac{k - 1}{Y + k - 1}. \tag{11.9}
\]

(That was not a proof. One needs to show directly that it is unbiased.)

Now we test out the inference: “The unbiased estimate of \( \theta \) is ...”:

- **Binomial(10, \theta)**, with outcome \( x = 4 \). “The unbiased estimate of \( \theta \) is \( \hat{\theta}_U = 4/10 = 2/5 \”

- **Negative Binomial(4, \theta)**, with outcome \( y = 6 \). “The unbiased estimate of \( \theta \) is \( \hat{\theta}_U = 3/9 = 1/3 \”

Those two situations have the same likelihood, but different estimates, thus violating the likelihood principle!\(^1\) On the other hand, the MLE in both cases would be 2/5.

The problem with unbiasedness is that it depends on the entire density, i.e., on outcomes not observed, so different densities would give different expected values. For that reason, any inference that involves the operating characteristics of the procedure violates the likelihood principle.

Whether one decides to fully accept the likelihood principle or not, it provides an important guide for any kind of inference, as we shall see.

### 11.2 Sufficiency

Consider again (11.3), or more generally, \( X_1, \ldots, X_n \) iid \( N(\mu, 1) \). The likelihood is

\[
L(\mu : x_1, \ldots, x_n) = e^{\mu \sum x_i - \frac{1}{2} \mu^2}, \tag{11.10}
\]

where the \( \exp(-1/2 \sum x_i^2) \) part can be dropped as it does not depend on \( \mu \). Note that this function depends on the \( x_i \)'s only through their sum, that is, as in (11.5), if \( \mathbf{x} \) and \( \mathbf{x}^* \) have the same sum, they have the same likelihood. Thus as far as the likelihood goes, all we need to know is the sum of the \( x_i \)'s to make an inference about \( \mu \). This sum is a sufficient statistic.

**Definition 17.** Consider the model with space \( \mathcal{X} \) and parameter space \( \Theta \). A function

\[
s : \mathcal{X} \rightarrow S \tag{11.11}
\]

is a sufficient statistic if for some function \( b \),

\[
b : S \times \Theta \rightarrow [0, \infty), \tag{11.12}
\]

the likelihood can be written

\[
L(\theta ; x) = b(s(x), \theta). \tag{11.13}
\]

\(^1\)Don’t worry too much. You won’t be arrested.
Thus $S = s(X)$ is a sufficient statistic (it may be a vector) if by knowing $S$, you know the likelihood, i.e., it is sufficient for performing any inference. That is handy, because you can reduce your data set, which may be large, to possibly just a few statistics without losing any information. More importantly, it turns out that the best inferences depend on just the sufficient statistics.

We next look at some examples.

11.2.1 Examples

First, we note that for any model, the data $X$ is itself sufficient, because the likelihood depends on $X$ through $X$.

IID

If $X_1, \ldots, X_n$ are iid with density $f(x_i; \theta)$, then no matter what the model, the order statistics (see Section 6.3) are sufficient. To see this fact, write

$$L(\theta; x) = f(x_1; \theta) \cdots f(x_n; \theta) = f(x_{(1)}; \theta) \cdots f(x_{(n)}; \theta) = b((x_{(1)}, \ldots, x_{(n)}), \theta),$$

(11.14)

because the order statistics are just the $x_i$’s in a particular order.

Normal

If $X_1, \ldots, X_n$ are iid $N(\mu, 1)$, $\Theta = \mathbb{R}$, then we have several candidates for sufficient statistic:

- The data itself: $s_1(x) = \bar{x}$
- The order statistics: $s_2(x) = (x_{(1)}, \ldots, x_{(n)})$
- The sum: $s_3(x) = \sum x_i$
- The mean: $s_4(x) = \bar{x}$
- Partial sums: $s_5(x) = (x_1 + x_2, x_3 + x_4 + x_5, x_6)$ (if $n = 6$).

An important fact is that any one-to-one function of a sufficient statistic is also sufficient, because knowing one means you know the other, hence you know the likelihood. For example, the mean and sum are one-to-one. Also, note that the dimension of the sufficient statistics in (11.15) are different ($n$, $n$, $1$, $1$, and $3$, respectively). Generally, one prefers the most compact one, in this case either the mean or sum. Each of those are functions of the others. In fact, they are minimal sufficient.

Definition 18. A statistic $s(x)$ is **minimal sufficient** if it is sufficient, and given any other sufficient statistic $t(x)$, there is a function $h$ such that $s(x) = h(t(x))$.

This concept is important, but we will focus on a more restrictive notion of completeness.

Which statistics are sufficient depends crucially on the parameter space. In the above, we assumed the variance known. But suppose $X_1, \ldots, X_N$ are iid $N(\mu, \sigma^2)$, with $(\mu, \sigma^2) \in \Theta \equiv \mathbb{R} \times (0, \infty)$. Then the likelihood is

$$L(\mu, \sigma^2; x) = \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2} = \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} \sum x_i^2 + \frac{1}{\sigma^2} \mu \sum x_i - \frac{1}{2\sigma^2} \mu^2}.$$

(11.16)

Now we cannot eliminate the $\sum x_i^2$ part, because it involves $\sigma^2$. Here the sufficient statistic is two-dimensional:

$$s(x) = (s_1(x), s_2(x)) = (\sum x_i, \sum x_i^2),$$

(11.17)
and the $b$ function is

$$b(s_1, s_2) = \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} s_2 + \frac{1}{2\sigma^2} \mu s_1 - \frac{1}{2\sigma^2} \mu^2}.$$  (11.18)

**Uniform**

Suppose $X_1, \ldots, X_n$ are iid $\text{Uniform}(0, \theta)$, $\theta \in (0, \infty)$. The likelihood is

$$L(\theta ; \mathbf{x}) = \prod \frac{1}{\theta} I_{\{0 < x_i < \theta\}}(\theta) = \left\{ \begin{array}{ll} \frac{1}{\theta} & \text{if } 0 < x_i < \theta \text{ for all } x_i \\ 0 & \text{if not} \end{array} \right.$$  (11.19)

All the $x_i$'s are less than $\theta$ if and only if the largest one is, so that we can write

$$L(\theta ; \mathbf{x}) = \left\{ \begin{array}{ll} \frac{1}{\theta} & \text{if } 0 < x_{(n)} < \theta \\ 0 & \text{if not} \end{array} \right.$$  (11.20)

Thus the likelihood depends on $\mathbf{x}$ only through the maximum, hence $x_{(n)}$ is sufficient.

**Double exponential**

Now suppose $X_1, \ldots, X_n$ are iid $\text{Double Exponential}(\theta)$, $\theta \in \mathbb{R}$, which has pdf $f(x_i | \theta) = (1/2) \exp(-|x_i - \theta|)$, so that the likelihood is

$$L(\theta ; \mathbf{x}) = e^{-\sum |x_i - \theta|}.$$  (11.21)

Because the data are iid, the order statistics are sufficient, but unfortunately, there is not another sufficient statistic with smaller dimension. The absolute value bars cannot be removed. Similar models based on the Cauchy, logistic, and others have the same problem.

**Exponential families**

In some cases, the sufficient statistic does reduce the dimensionality of the data significantly, such as in the iid normal case, where no matter how large $n$ is, the sufficient statistic is two-dimensional. In other cases, such as the double exponential above, there is no dimensionality reduction, so one must still carry around $n$ values. Exponential families are special families in which there is substantial reduction for iid variables or vectors. Because the likelihood of an iid sample is the product of individual likelihoods, statistics “add up” when they are in an exponent.

The vector $\mathbf{X}$ has an **exponential family** distribution if its density (pdf or pmf) depends on a $p \times 1$ vector $\underline{\theta}$ that can be written

$$f(\mathbf{x} | \underline{\theta}) = a(\mathbf{x}) e^{\theta_1 t_1(\mathbf{x}) + \cdots + \theta_p t_p(\mathbf{x}) - \psi(\underline{\theta})}$$  (11.22)

for some functions $t_1(\mathbf{x}), \ldots, t_p(\mathbf{x})$, $a(\mathbf{x})$, and $\psi(\underline{\theta})$. The $\underline{\theta}$ is called the **natural parameter** and the $t(\mathbf{x}) = (t_1(\mathbf{x}), \ldots, t_p(\mathbf{x}))'$ is the vector of **natural sufficient statistics**. Now if $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are iid vectors with density (11.22), then the joint density is

$$f(\mathbf{X}_1, \ldots, \mathbf{X}_n | \underline{\theta}) = \prod f(\mathbf{x}_i | \underline{\theta}) = [\prod a(\mathbf{x}_i)] e^{\theta_1 \sum t_1(\mathbf{x}) + \cdots + \theta_p \sum t_p(\mathbf{x}) - n \psi(\underline{\theta})},$$  (11.23)
hence has sufficient statistic

\[
s(x_1, \ldots, x_n) = \left( \frac{\sum_i t_1(x_i)}{\sum_i t_p(x_i)} \right) = \sum_i t(x_i),
\]
which has dimension \( p \) no matter how large \( n \).

The natural parameters and sufficient statistics are not necessarily the most "natural" to us. For example, in the normal case of (11.16), the natural parameters can be taken to be

\[
\theta_1 = \frac{\mu}{\sigma^2} \quad \text{and} \quad \theta_2 = -\frac{1}{2\sigma^2}.
\]

The corresponding statistics are

\[
t(x_i) = \left( \frac{t_1(x_i)}{t_2(x_i)} \right) = \left( \frac{x_i}{\bar{x}_i^2} \right).
\]

There are other choices, e.g., we could switch the negative sign from \( \theta_2 \) to \( t_2 \).

Other examples are the Poisson(\( \lambda \)), which has \( \theta = \log(\lambda) \), and the Binomial(\( n, p \)), which has \( \theta = \log(p/(1-p)) \); also the gamma, beta, multivariate normal, multinomial, and others.

### 11.3 Conditioning on a sufficient statistic

The intuitive meaning of a sufficient statistic is that once you know the statistic, nothing else about the data helps in inference about the parameter. For example, in the iid case, once you know the values of the \( x_i \)'s, it does not matter in what order they are listed. This notion can be formalized by finding the conditional distribution of the data given the sufficient statistic, and showing that it does not depend on the parameter.

First, a lemma that makes it easy to find the conditional density, if there is one.

**Lemma 13.** Suppose \( X \) has space \( \mathcal{X} \) and density \( f_X \), and \( S = s(X) \) is a function of \( X \) with space \( \mathcal{S} \) and density \( f_S \). Then the conditional density of \( X \) given \( S \), if it exists, is

\[
f_{X|S}(x | s) = \frac{f_X(x)}{f_S(s)} \quad \text{for} \quad x \in \mathcal{X}_s \equiv \{ x \in \mathcal{X} \mid s(x) = s \}.
\]

The caveat "if it exists" in the lemma is unnecessary in the discrete case, because the conditional pmf always will exist. But in continuous or mixed cases, the resulting conditional distribution may not have a density with respect to Lebesgue measure. It will have a density with respect to some measure, though, which one would see in a measure theory course.

**Proof, at least in the discrete case.** Suppose \( X \) is discrete. Then by Bayes Theorem,

\[
f_{X|S}(x | s) = \frac{f_{X|S}(s | x) f_X(x)}{f_S(s)}, \quad x \in \mathcal{X}_s.
\]
Because $s$ is a function of $X$,

$$f_{S|X} (s \mid x) = P[s(X) = s \mid X = x] = \begin{cases} 1 & \text{if } s(x) = s \\ 0 & \text{if } s(x) \neq s \end{cases} = \begin{cases} 1 & \text{if } x \in X_s \\ 0 & \text{if } x \notin X_s \end{cases}$$

Thus the term "$f_{S|X} (s \mid x)$" equals 1 in (11.28), because $x \in X_S$, so we can erase it, yielding (11.27).

Here is the main result.

**Lemma 14.** Suppose $s(x)$ is a sufficient statistic for a model with data $X$ and parameter space $\Theta$. Then the conditional distribution $X \mid s(X) = s$ does not depend on $\theta$.

Before giving a proof, consider the example with $X_1, \ldots, X_n$ iid Poisson($\theta$), $\theta \in (0, \infty)$. The likelihood is

$$L(\theta ; x) = \prod e^{-\theta} \theta^{x_i} = e^{-n\theta} \theta^{\sum x_i}.$$  \hfill (11.30)

We see that $s(x) = \sum x_i$ is sufficient. We know that $S = \sum X_i$ is Poisson($n\theta$), hence

$$f_S(s) = e^{-n\theta} \frac{(n\theta)^s}{s!}.$$ \hfill (11.31)

Then by Lemma 13, the conditional pmf of $X$ given $S$ is

$$f_{X|S} (x \mid s) = \frac{f_X(x)}{f_S(s)} = \frac{e^{-n\theta} \theta^{\sum x_i} / \prod x_i!}{e^{-n\theta} (n\theta)^s / s!} = \frac{s!}{\prod x_i!} \frac{1}{n^s} \frac{1}{s!} x \in X_s = \{ x \mid \sum x_i = s \}. \hfill (11.32)$$

That is a multinomial distribution:

$$X \mid \sum X_i = s \sim \text{Multinomial}_n (s, \frac{1}{n}, \ldots, \frac{1}{n}).$$  \hfill (11.33)

But the main point of Lemma 14 is that this distribution is independent of $\theta$. Thus knowing the sum means the exact values of the $x_i$'s do not reveal anything extra about $\theta$.

The key to the result lies in (11.30) and (11.31), where we see that the likelihoods of $X$ and $s(X)$ are the same, if $\sum x_i = s$. This fact is a general one.

**Lemma 15.** Suppose $s(X)$ is a sufficient statistic for the model with data $X$, and consider the model for $S$ where $S = D s(X)$. Then the likelihoods for $X = x$ and $S = s$ are the same if $s = s(x)$.

This result should make sense, because it is saying that the sufficient statistic contains the same information about $\theta$ that the full data do. One consequence is that instead of having to work with the full model, one can work with the sufficient statistic's model. For example, instead of working with $X_1, \ldots, X_n$ iid $N(\mu, \sigma)$, you can work
with just the two independent random variables $\bar{X} \sim N(\mu, \sigma^2 / n)$ and $S^2 \sim \sigma^2 \chi^2_{n-1} / n$ without losing any information.

We prove the lemma in the discrete case.

**Proof.** Let $f_X(x | \theta)$ be the pmf of $X$, and $f_S(s | \theta)$ be that of $S$. Because $s(X)$ is sufficient, the likelihood for $X$ can be written

$$L(\theta; x) = c_x f_X(x | \theta) = b(s(x), \theta)$$

(11.34)

by Definition 17, for some $c_x$ and $b(s, \theta)$. The pmf of $S$ is

$$f_S(s | \theta) = p_\theta[s(X) = s] = \sum_{x \in X_s} f_X(x | \theta) \quad (\text{where } X_s = \{ x \in X | s(x) = s \})$$

$$= \sum_{x \in X_s} b(s(x), \theta) / c_x$$

$$= b(s, \theta) \sum_{x \in X_s} 1 / c_x \quad (\text{because in the summation, } s(x) = s)$$

$$= b(s, \theta) d_s$$

(11.35)

where $d_s$ is that sum of $1 / c_x$'s. Thus the likelihood of $S$ can be written

$$L^*(\theta; s) = b(s, \theta),$$

(11.36)

the same as $L$ in (11.34).

A formal proof for the continuous case proceeds by introducing appropriate extra variables $Y$ so that $X$ and $(S, Y)$ are one-to-one, then using Jacobians, then integrating out the $Y$. We will not do that, but special cases can be done easily, e.g., if $X_1, \ldots, X_n$ are iid $N(\mu, 1)$, one can directly show that $X$ and $s(X) = \sum X_i \sim N(n\mu, n)$ have the same likelihood.

Now for the proof of Lemma 14, again in the discrete case. Basically, we just repeat the calculations for the Poisson.

**Proof.** As in the proof of Lemma 15, the pmf's of $X$ and $S$ can be written, respectively,

$$f_X(x | \theta) = L(\theta; x) / c_x \quad \text{and} \quad f_S(s | \theta) = L^*(\theta; s) d_s$$

(11.37)

where the likelihoods are equal in the sense that

$$L(\theta; x) = L^*(\theta; s) \quad \text{if} \quad x \in X_s.$$  

(11.38)

Then by Lemma 13,

$$f_{X|S}(x | s, \theta) = \frac{f_X(x | \theta)}{f_S(s | \theta)} \quad \text{for} \quad x \in X_s,$$

$$= \frac{L(\theta; x) / c_x}{L^*(\theta; s) d_s} \quad \text{for} \quad x \in X_s,$$

$$= \frac{1}{c_x d_s} \quad \text{for} \quad x \in X_s,$$

(11.39)

by (11.38), which does not depend on $\theta$.

We end this section with a couple of additional examples that derive the conditional distribution of $X$ given $s(X)$.  

\]
11.3. Conditioning on a sufficient statistic

11.3.1 Example: IID

Suppose \( X_1, \ldots, X_n \) are iid continuous random variables with pdf \( f(x_i) \). No matter what the model, we know that the order statistics are sufficient, so we will suppress the \( \theta \). We can proceed as in the proof of Lemma 14. Letting \( s(\mathbf{x}) = (x_{(1)}, \ldots, x_{(n)}) \), we know that the joint pdf’s of \( \mathbf{X} \) and \( \mathbf{s} \equiv s(\mathbf{X}) \) are, respectively,

\[
 f_x(x) = \prod f(x_i) \quad \text{and} \quad f_s(s) = n! \prod f(x_{(i)}).
\] (11.40)

The products of the pdf’s are the same, just written in different orders. Thus

\[
 P[\mathbf{X} = \mathbf{x} \mid s(\mathbf{X}) = \mathbf{s}] = \frac{1}{n!} \quad \text{for} \quad \mathbf{x} \in \mathcal{X}_s.
\] (11.41)

The \( s \) is a particular set of ordered values, and \( \mathcal{X}_s \) is the set of all \( \mathbf{x} \) that have the same values as \( s \), but in any order. To illustrate, suppose that \( n = 3 \) and \( s(\mathbf{X}) = s = (1, 2, 7) \). Then \( \mathbf{X} \) has a conditional chance of \( 1/6 \) of being any \( \mathbf{x} \) with order statistics \( (1, 2, 7) \):

\[
 \mathcal{X}_{(1,2,7)} = \{(1, 2, 7), (1, 7, 2), (2, 1, 7),
\]

\[
 (2, 7, 1), (7, 1, 2), (7, 2, 1)\}. \quad (11.42)
\]

The discrete case works as well, although the counting is a little more complicated when there are ties. For example, if \( s(\mathbf{x}) = (1, 3, 3, 4) \), there are \( 4!/2! = 12 \) different orderings.

11.3.2 Example: Normal mean

Suppose \( X_1, \ldots, X_n \) are iid \( N(\mu, 1) \), so that \( \overline{X} \) is sufficient (as is \( \sum X_i \)). We wish to find the conditional distribution of \( \mathbf{X} \) given \( \overline{X} = \mathbf{x} \). Because we have normality and linear functions, everything is straightforward. First we need the joint distribution of \( \mathbf{W} = (\overline{X}, \mathbf{X}')' \), which is a linear transformation of \( \mathbf{X} \sim N(\mu \mathbf{1}_n, \mathbf{I}_n) \), hence multivariate normal. We could figure out the matrix for the transformation, but all we really need are the mean and covariance matrix of \( \mathbf{W} \). The mean and covariance of the \( \mathbf{X} \) part we know, and the mean and variance of \( \overline{X} \) are \( \mu \) and \( 1/n \), respectively. All that is left is the covariance of the \( X_i \)'s with \( \overline{X} \), which are all the same, and can be found directly:

\[
 \text{Cov}[X_i, \overline{X}] = \frac{1}{n} \sum_{j=1}^{n} \text{Cov}[X_i, X_j] = \frac{1}{n},
\] (11.43)

because \( X_i \) is independent of all the \( X_j \)'s except for \( X_i \) itself, with which its covariance is 1. Thus

\[
 \mathbf{W} = \left( \begin{array}{c} \overline{X} \\ \mathbf{X} \end{array} \right) \sim N \left( \mu \mathbf{1}_n + \mathbf{1}_n, \left( \begin{array}{cc} \frac{1}{n} \mathbf{1}_{n+1} & \frac{1}{n} \mathbf{1}_n \\ \frac{1}{n} \mathbf{1}_n' & \mathbf{I}_n \end{array} \right) \right).
\] (11.44)

For the conditional distribution, we use (8.101) through (8.103). The \( X \) there is \( \overline{X} \) here, and the \( Y \) there is the \( \mathbf{X} \) here. Thus

\[
 \Sigma_{XX} = 1/n, \quad \Sigma_{XY} = (1/n)\mathbf{1}_n, \quad \Sigma_{YY} = \mathbf{I}_n, \quad \mu_X = \mu, \quad \text{and} \quad \mu_Y = \mu \mathbf{1}_n, \quad (11.45)
\]

and

\[
 \beta = (1/n)\mathbf{1}_n (1/n)^{-1} = \mathbf{1}_n \quad \text{and} \quad \alpha = \mu \mathbf{1}_n - \beta \mu = \mu \mathbf{1}_n - \mathbf{1}_n \mu = \mathbf{0}_n. \quad (11.46)
\]
Thus the conditional mean is
\[ E[X \mid X = \bar{x}] = \alpha + \beta \bar{x} = \bar{x} \mathbb{1}_n. \] (11.47)

That result should not be surprising. It says that if you know the sample mean is \( \bar{x} \), you expect on average the observations to be \( \bar{x} \). The conditional covariance is
\[
\Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} = I_n - (1/n) \mathbb{1}_n (1/n)^{-1} (1/n) \mathbb{1}_n' = I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n' = H_n, \tag{11.48}
\]
the centering matrix from (8.42). Putting it all together:
\[ X \mid X = \bar{x} \sim N(\bar{x} \mathbb{1}_n, H_n). \tag{11.49} \]

We can pause and reflect that this distribution is free of \( \mu \), as it had better be. Note also that if we subtract the conditional mean, we have
\[ X - \bar{x} \mathbb{1}_n \mid \bar{x} = \bar{x} \sim N(0_n, H_n). \tag{11.50} \]

That conditional distribution is free of \( \bar{x} \), meaning the vector is independent of \( \bar{x} \). But this is the vector of deviations, and we already knew it is independent of \( \bar{x} \) from (8.46).

### 11.4 Rao-Blackwell: Improving an estimator

We see that sufficient statistics are nice in that we do not lose anything by restricting to them. In fact, they are more than just convenient ... if you base an estimate on more of the data than the sufficient statistic, then you do lose something. For example, suppose \( X_1, \ldots, X_n \) are iid \( N(\mu, 1) \), \( \mu \in \mathbb{R} \), and we wish to estimate
\[ g(\mu) = P_{\mu}[X_i \leq 10] = \Phi(10 - \mu), \tag{11.51} \]
\( \Phi \) being the distribution function of the standard normal. An unbiased estimator is
\[ \delta(\bar{x}) = \frac{\#\{x_i \mid x_i \leq 10\}}{n} = \frac{1}{n} \sum I_{\{x_i \leq 10\}}(x_i). \tag{11.52} \]

Note that this estimator is not a function of just \( \bar{x} \), the sufficient statistic. The claim is that there is then another unbiased estimator which is a function of just \( \bar{x} \) that is better, i.e., has lower variance.

Fortunately, there is a way to find such an estimator besides guessing. We use conditioning as in the previous section, and the results on conditional means and variances in Section 7.4.3. Start with an estimator \( \delta(\bar{x}) \) of \( g(\theta) \), and a sufficient statistic \( S = s(\bar{x}) \). Then consider the conditional expected value of \( \delta \):
\[ \delta^* = E[\delta(\bar{x}) \mid s(\bar{x}) = \bar{s}]. \tag{11.53} \]

First, we need to make sure \( \delta^* \) is an estimator, which means that it does not depend on the unknown \( \theta \). But by Lemma 14, we know that the conditional distribution of \( \bar{x} \) given \( \bar{s} \) does not depend on \( \theta \), and \( \delta \) does not depend on \( \theta \) because it is an estimator, hence \( \delta^* \) is an estimator. If we condition on something not sufficient, then we may not end up with an estimator.
11.4. Rao-Blackwell: Improving an estimator

Is \( \delta^* \) a good estimator? From (7.43), we know it has the same expected value as \( \delta \):

\[
E[\delta^*(S) \mid \theta] = E[\delta(X) \mid \theta],
\]
so that

\[
\text{Bias}(\theta; \delta^*) = \text{Bias}(\theta; \delta).
\]

Thus we haven’t done worse in terms of bias, and in particular if \( \delta \) is unbiased, so is \( \delta^* \).

Turning to variance, we have the equation (7.48), which translated to our situation here is

\[
\text{Var}[\delta(X) \mid \theta] = E[v(S) \mid \theta] + \text{Var}[\delta^*(S) \mid \theta],
\]

where

\[
v(s) = \text{Var}[\delta(X) \mid s(X) = s, \theta].
\]

Whatever \( v \) is, it is not negative, hence

\[
\text{Var}[\delta^*(S) \mid \theta] \leq \text{Var}[\delta(X) \mid \theta].
\]

Thus variance-wise, the \( \delta^* \) is no worse than \( \delta \). In fact, \( \delta^* \) is strictly better unless \( v(S) \) is zero with probability one, but in that case, then \( \delta \) and \( \delta^* \) are the same, so that \( \delta \) itself is a function of just \( S \) already.

Finally, if the bias is the same and the variance of \( \delta^* \) is lower, then the mean squared error of \( \delta^* \) is better. To summarize:

**Theorem 7. Rao-Blackwell**. If \( \delta \) is an estimator of \( g(\theta) \), and \( s(X) \) is sufficient, then \( \delta^*(S) \) given in (11.53) has the same bias as \( \delta \), and smaller variance and MSE, unless \( \delta \) is a function of just \( S \), in which case \( \delta \) and \( \delta^* \) are the same.

11.4.1 Example: Normal probability

Consider the estimator \( \delta(x) \) in (11.52) for \( g(\mu) = \Phi(10 - \mu) \) in (11.51) in the normal case. With \( \overline{X} \) being the sufficient statistic, we can find the conditional expected value of \( \delta \). We start by finding the conditional expected value of just one of the \( I_{\{X_i \leq 10\}}(x_i)'s \). It turns out that the conditional expectation is the same for each \( i \), hence the conditional expected value of \( \delta \) is the same as the condition expected value of just one. So we are interested in finding

\[
\delta^*(\overline{X}) = E[I_{\{X_i \leq 10\}}(X_i) \mid \overline{X} = \overline{x}] = P[X_i \leq 10 \mid \overline{X} = \overline{x}].
\]

From (11.49) we have that

\[
X_i \mid \overline{X} = \overline{x} \sim N(\overline{x}, 1 - \frac{1}{n}),
\]

hence

\[
\delta^*(\overline{x}) = P[X_i \leq 10 \mid \overline{X} = \overline{x}]
\]

\[= P[N(\overline{x}, 1 - 1/n) \leq 10]
\]

\[= \Phi \left( \frac{10 - \overline{x}}{\sqrt{1 - 1/n}} \right)
\]

\[= \Phi \left( \sqrt{\frac{n - 1}{n}} (10 - \overline{x}) \right).
\]
This estimator is then guaranteed to be unbiased, and have a lower variance that \( \delta \). It would have been difficult to come up with this estimator directly, or even show that it is unbiased, but the original \( \delta \) is quite straightforward, as is the conditional calculation.

### 11.4.2 Example: IID

In Example 11.3.1, we saw that when the observations are iid, the conditional distribution of \( \overline{X} \) given the order statistics is uniform over all the permutations of the observations. One consequence is that any estimator must be invariant under permutations. For a simple example, consider estimating \( \mu \), the mean, with the simple estimator \( \delta(X) = X_1 \). Then with \( s(x) \) being the order statistics,

\[
\delta^*(s) = E[X_1 \mid s(X) = s] = \frac{1}{n} \sum s_i = \overline{s} \tag{11.62}
\]

because \( X_1 \) is conditionally equally likely to be any of the order statistics. Of course, the mean of the order statistics is the same as the \( \overline{x} \), so we have that \( \overline{X} \) is a better estimate than \( X_1 \), which we knew. The procedure applied to any weighted average will also end up with the mean, e.g.,

\[
E \left[ \frac{1}{2} X_1 + \frac{1}{3} X_2 + \frac{1}{6} X_4 \mid s(X) = \overline{s} \right] = \frac{1}{2} E[X_1 \mid s(X) = \overline{s}] + \frac{1}{3} E[X_2 \mid s(X) = \overline{s}] + \frac{1}{6} E[X_4 \mid s(X) = \overline{s}] \\
= \frac{1}{2} \overline{s} + \frac{1}{3} \overline{s} + \frac{1}{6} \overline{s} \\
= \overline{s}. \tag{11.63}
\]

Turning to the variance \( \sigma^2 \), because \( X_1 - X_2 \) has mean 0 and variance \( 2\sigma^2 \), \( \delta(x) = (x_1 - x_2)^2/2 \) is an unbiased estimator of \( \sigma^2 \). Conditioning on the order statistics, we obtain the mean of all the \( (x_i - x_j)^2/2 \)'s with \( i \neq j \):

\[
\delta^*(s) = E \left[ \frac{(X_1 - X_2)^2}{2} \right] = \frac{1}{n(n-1)} \sum \sum_{i \neq j} \frac{(x_i - x_j)^2}{2}. \tag{11.64}
\]

After some algebra, we can write

\[
\delta^*(s(x)) = \frac{\sum (x_i - \overline{x})^2}{n-1}, \tag{11.65}
\]

the usual unbiased estimate of \( \sigma^2 \).

The above estimators are special cases of \textbf{U-statistics}, for which there are many nice asymptotic results. They are studied in STAT575. A U-statistic is based on a kernel \( h(x_1, \ldots, x_d) \), a function of a subset of the observations. The corresponding U-statistic is the symmetrized version of the kernel, i.e., the conditional expected value,

\[
u(\overline{x}) = E[h(X_1, \ldots, X_d) \mid s(X) = s(\overline{x})] = \frac{1}{n(n-1) \cdots (n-d+1)} \sum \sum_{i_1, \ldots, i_d, \text{distinct}} h(x_{i_1}, \ldots, x_{i_d}). \tag{11.66}
\]
Chapter 12

UMVUE’s

12.1 Unbiased estimation

Section 10.6 introduces the notion of unbiased estimators. Several questions arise when estimating a particular parameter in a particular model:

• Are there any unbiased estimators?
• Is there a unique unbiased estimator?
• If there is no unique one, which (if any) is best?
• How does one find unbiased estimators?

The answer to all of the questions is, “It depends.” In particular, there is no general automatic method for finding unbiased estimators, in contrast to maximum likelihood estimators, or Bayes estimators. The latter two types may be difficult to calculate, but it is a mathematics or computational problem, not a statistical one.

One method that occasionally works for finding unbiased estimators is to find a power series from the definition. For example, suppose \( X \sim \text{Binomial}(n, \theta) \), \( \theta \in (0, 1) \). We know \( X/n \) is an unbiased estimator for \( \theta \), but what about estimating \( g(\theta) = \theta(1 - \theta) \), the numerator of the variance? An unbiased \( \delta \) will satisfy

\[
E_\theta[\delta(X)] = \theta(1 - \theta) \quad \text{for all } \theta \in (0, 1),
\]

or

\[
\sum_{x=0}^{n} \delta(x) \binom{n}{x} \theta^x (1-\theta)^{n-x} = \theta(1-\theta).
\]

Both sides of equation (12.2) are polynomials in \( \theta \), hence for equality to hold for every \( \theta \in (0, 1) \), the coefficients of each power \( \theta^k \) must match.

First, suppose \( n = 1 \), so that we need

\[
\delta(0)(1-\theta) + \delta(1)\theta = \theta(1-\theta) = \theta - \theta^2.
\]

There is no \( \delta \) that will work, because the coefficient of \( \theta^2 \) on the right-hand side is -1, and on the left-hand side is 0. That is, with \( n = 1 \), there is no unbiased estimator of \( \theta(1 - \theta) \).
Chapter 12. UMVUE’s

With \( n = 2 \), we have

\[
\begin{align*}
\delta(0)(1 - \theta)^2 + 2\delta(1)\theta(1 - \theta) + \delta(2)\theta^2 & = \theta(1 - \theta) \\
\delta(0) + (-2\delta(0) + 2\delta(1))\theta + (\delta(0) - 2\delta(1) + \delta(2))\theta^2 & = \theta - \theta^2.
\end{align*}
\]

(12.4)

Matching coefficients of \( \theta^k \):

<table>
<thead>
<tr>
<th>( k )</th>
<th>Left-hand side</th>
<th>Right-hand side</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( \delta(0) )</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>(-2\delta(0) + 2\delta(1))</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>( \delta(0) - 2\delta(1) + \delta(2) )</td>
<td>-1</td>
</tr>
</tbody>
</table>

(12.5)

we see that the only solution is

\[
\begin{align*}
\delta(0) & = 0 \\
\delta(1) & = \frac{1}{2} \\
\delta(2) & = 0,
\end{align*}
\]

(12.6)

which actually is easy to see directly from the first line of (12.4). Thus the \( \delta \) in (12.6) must be the best unbiased estimator, being the only one. In fact, because for any function \( \delta(x) \), \( E_\theta[\delta(X)] \) is an \( n^{th} \)-degree polynomial in \( \theta \), the only functions \( g(\theta) \) that have unbiased estimators are those that are themselves polynomials of degree \( n \) or less. For example, \( 1/\theta \) and \( e^\theta \) do not have unbiased estimators.

For \( n > 2 \), estimating \( \theta(1 - \theta) \) is more complicated, but in principle will yield a best unbiased estimator. Or we can use the approach conditioning on a sufficient statistic in Section 11.4, particularly Example 11.4.2. We know that \( X = Z_1 + \ldots + Z_n \), where the \( Z_i \)'s are iid Bernoulli(\( \theta \)). The estimator \( \delta \) in (12.6) for \( n = 2 \) Bernoulli’s can be given as \( (z_1 - z_2)^2/2 \). (Check it.) Then conditioning on the order statistics of the \( Z_i \)'s we obtaining the same result as in (11.65), i.e., a better unbiased estimator of \( \theta(1 - \theta) \) is

\[
\frac{\sum(Z_i - \bar{Z})^2}{n - 1} = \frac{\sum Z_i^2 - (\sum Z_i)^2/n}{n - 1} = \frac{\sum Z_i - (\sum Z_i)^2/n}{n - 1} \quad \text{(because } Z_i^2 = Z_i) \\
= \frac{X - (X)^2/n}{n - 1} \quad \text{(because } X = \sum Z_i) \\
= \frac{X(1 - \frac{1}{n}X)}{n - 1}.
\]

(12.7)

Is this estimator is the only unbiased one (of \( \theta(1 - \theta) \)) for the binomial? Yes, which can be shown using the results in Section 12.2. (Or if we had continued matching coefficients, we would have found it to be unique.)

Another approach is to take a biased estimator, and see if it can be modified to become unbiased. For example, if \( X_1, \ldots, X_n \) are iid Poisson(\( \theta \)), the MLE of \( g(\theta) = e^{-\theta} \) is

\[
\hat{g}(\theta) = e^{-X}.
\]

(12.8)
Because $S = \sum X_i \sim \text{Poisson}(\theta)$,

$$E[\hat{g}(\theta) | \theta] = E[e^{-S/n} | \theta] = \sum e^{-n\theta n^s} \frac{n^s}{s!} e^{-s/n} = e^{-n\theta} \sum \frac{(n\theta e^{-1/n})^s}{s!} = e^{-n\theta} e^{n\theta e^{-1/n}} \neq e^{-\theta}. \quad (12.9)$$

So the MLE is biased. Now $e^{-1/n} \approx (1 - 1/n)$. and if we make that substitution in the exponent of $E[\hat{g}(\theta)]$, it would be unbiased. Thus if we modified the $S/n$ a bit we may find an unbiased estimator. Consider

$$\hat{g}(\theta)_c = e^{-c\bar{X}} \quad (12.10)$$

for some constant $c$. Working through (12.9) again but with $c/n$ in place of $1/n$, we obtain

$$E[\hat{g}(\theta)_c | \theta] = e^{-n\theta} e^{n\theta e^{-c/n}}. \quad (12.11)$$

We want

$$-n + ne^{-c/n} = -1 \Rightarrow c = -n \log(1 - 1/n) \equiv c^*, \quad (12.12)$$

hence an unbiased estimator is

$$\hat{g}(\theta)_{c^*} = \left(1 - \frac{1}{n}\right)^{\bar{X}}. \quad (12.13)$$

We have seen that there may or may not be an unbiased estimator. Often, there are many. For example, suppose $X_1, \ldots, X_n$ are iid $\text{Poisson}(\theta)$. Then $\bar{X}$ and $\sum(X_i - \bar{X})^2/(n-1)$ are both unbiased, being unbiased estimators of the mean and variance, respectively. Weighted averages of the $X_i$’s are also unbiased, e.g.,

$$X_1, \frac{X_1 + X_2}{2}, \text{ and } \frac{1}{3} X_1 + \frac{1}{2} X_2 + \frac{1}{6} X_3 \quad (12.14)$$

are all unbiased. Among all the possibilities, is there one that is best, i.e., has the lowest variance? There is not a good general answer, but the rest of this chapter uses sufficiency and the concept of completeness to find unique best unbiased estimators in many situations.

### 12.2 Completeness: Unique unbiased estimators

We have already answered the question of finding the best unbiased estimator in certain situations without realizing it. We know that any estimator that is not a function of just the sufficient statistic can be improved. Also, if there is only one unbiased estimator that is a function of the sufficient statistic, then it must be the best one that depends on only the sufficient statistic. Furthermore, it must be the best overall, because it is better than any estimator that is not a function of just the sufficient statistic.
Chapter 12. UMVUE’s

The concept of “there being only one unbiased estimator” is called completeness. It is a property attached to a model. Consider the model with random vector $T$, space $T$, and parameter space $\Theta$. Suppose there are at least two unbiased estimators of some function $g(\theta)$ in this model, say $\delta_{g}(T)$ and $\delta^*_g(T)$. That is,

\[ P_{\theta}[\delta_{g}(T) \neq \delta^*_g(T)] > 0 \text{ for some } \theta \in \Theta, \quad (12.15) \]

and

\[ E_{\theta}[\delta_{g}(T)] = g(\theta) = E_{\theta}[\delta^*_g(T)] \text{ for all } \theta \in \Theta. \quad (12.16) \]

Then

\[ E_{\theta}[\delta_{g}(T) - \delta^*_g(T)] = 0 \text{ for all } \theta \in \Theta. \quad (12.17) \]

This $\delta_{g}(T) - \delta^*_g(T)$ is an unbiased estimator of 0. Now suppose $\delta_{h}(T)$ is an unbiased estimator of the function $h(\theta)$. Then so is $\delta^*_h(T) = \delta_{h}(T) + \delta_{g}(T) - \delta^*_g(T)$,

\[ \delta^*_h(T) = \delta_{h}(T) + \delta_{g}(T) - \delta^*_g(T), \quad (12.18) \]

because

\[ E_{\theta}[\delta^*_h(T)] = E_{\theta}[\delta_{h}(T)] + E_{\theta}[\delta_{g}(T) - \delta^*_g(T)] = h(\theta) + 0. \quad (12.19) \]

That is, if there is more than one unbiased estimator of one function, then there is more than one unbiased estimator of any other function (that has at least one unbiased estimator).

Logically, it follows that if there is only one unbiased estimator of some function, then there is only one (or zero) unbiased estimator of any function. That one function may as well be the zero function.

**Definition 19.** Suppose for the model given by $(T, T, \Theta)$, the only unbiased estimator of 0 is 0 itself. That is, suppose

\[ E_{\theta}[\delta(T)] = 0 \text{ for all } \theta \in \Theta \quad (12.20) \]

implies that

\[ P_{\theta}[\delta(T) = 0] = 1 \text{ for all } \theta \in \Theta. \quad (12.21) \]

Then the model is complete.

The important implication follows.

**Lemma 16.** Suppose the model given by $(T, T, \Theta)$ is complete. Then there exists at most one unbiased estimator of any function $g(\theta)$.

Illustrating with the Binomial again, suppose $X \sim \text{Binomial}(n, \theta)$ with $\theta \in (0, 1)$. Suppose $\delta(x)$ is an unbiased estimator of 0:

\[ E_{\theta}[\delta(X)] = 0 \text{ for all } \theta \in (0, 1). \quad (12.22) \]

We know the left-hand side is a polynomial in $\theta$, as is the right-hand side. All the coefficients of $\theta^i$ are zero on the right, hence on the left. Write

\[ E_{\theta}[\delta(X)] = \delta(0) \binom{n}{0} (1 - \theta)^n + \delta(1) \binom{n}{1} \theta(1 - \theta)^{n-1} \]

\[ + \cdots + \binom{n}{n-1} \theta^{n-1}(1 - \theta) + \binom{n}{n} \theta^n. \quad (12.23) \]
The coefficient of $\theta^0$, i.e., the constant, arises from just the first term, so is $\delta(0)(\binom{n}{0})$. For that to be 0, we have $\delta(0) = 0$. Erasing that first term, we see that the coefficient of $\theta$ is $\delta(1)(1)$, hence $\delta(1) = 0$. Continuing, we see that $\delta(2) = \cdots = \delta(n) = 0$, which means that $\delta(x) = 0$, which means that the only unbiased estimator of 0 is 0 itself.

Hence this model is complete.

### 12.2.1 Example: Poisson

Suppose $X_1, \ldots, X_n$ are iid Poisson($\theta$), $\theta \in (0, \infty)$, with $n > 1$. Is this model complete? No. Consider $\delta(X) = x_1 - x_2$.

$$E_\theta[\delta(X)] = E_\theta[X_1] - E_\theta[X_2] = \theta - \theta = 0,$$

(12.25) but

$$P_\theta[\delta(X) = 0] = P_\theta[X_1 = X_2] > 0.$$  

(12.26)

So 0 is not the only unbiased estimator of 0; $X_1 - X_2$ is another. You can come up with an infinite number, in fact.

Now let $S = X_1 + \cdots + X_n$, which is a sufficient statistic. The model for $S$ then has space $S = \{0, 1, 2, \ldots\}$ and distribution $S \sim \text{Poisson}(n\theta)$ for $\theta \in (0, \infty)$. Is the model for $S$ complete? Suppose $\delta^*(s)$ is an unbiased estimator of 0. Then

$$E_\theta[\delta^*(S)] = 0 \quad \text{for all } \theta \in (0, \infty) \Rightarrow \sum_{s=0}^{\infty} \delta^*(s)e^{-n\theta}(n\theta)^s/s! = 0 \quad \text{for all } \theta \in (0, \infty)$$

$$\Rightarrow \sum_{s=0}^{\infty} \delta^*(s)\frac{n^s}{s!}\theta^s = 0 \quad \text{for all } \theta \in (0, \infty)$$

$$\Rightarrow \delta^*(s)\frac{n^s}{s!} = 0 \quad \text{for all } s \in S$$

$$\Rightarrow \delta^*(s) = 0 \quad \text{for all } s \in S.$$  

(12.27)

Thus the only unbiased estimator of 0 that is a function of $S$ is 0, meaning the model for $S$ is complete.

### 12.2.2 Example: Uniform

Suppose $X_1, \ldots, X_n$ are iid Uniform(0, $\theta$), $\theta \in (0, \infty)$, with $n > 1$. This model again is not complete. No iid model is complete, in fact, if $n > 1$. Consider the sufficient statistic $S = \max\{X_1, \ldots, X_n\}$. The model for $S$ has space $(0, \infty)$ and pdf

$$f_\theta(s) = \begin{cases} ns^{n-1}/\theta^n & \text{if } 0 < s < \theta \smallskip \\ 0 & \text{if not} \end{cases}.$$  

(12.28)

To see if the model for $S$ is complete, suppose that $\delta^*$ is an unbiased estimator of 0. Then

$$E_\theta[\delta^*(S)] = 0 \quad \text{for all } \theta \in (0, \infty) \Rightarrow \int_0^\theta \delta^*(s)s^{n-1}ds/\theta^n = 0 \quad \text{for all } \theta \in (0, \infty)$$

$$\Rightarrow \int_0^\theta \delta^*(s)s^{n-1}ds = 0 \quad \text{for all } \theta \in (0, \infty)$$

(taking $d/d\theta$)

$$\Rightarrow \delta^*(\theta)\theta^{n-1} = 0 \quad \text{for (almost) all } \theta \in (0, \infty)$$

$$\Rightarrow \delta^*(\theta) = 0 \quad \text{for (almost) all } \theta \in (0, \infty).$$  

(12.29)
Chapter 12. U MVUE’s

That is, $\delta^*$ must be 0, so that the model for $S$ is complete. [The “(almost)” means that one can deviate from zero for a few values (with total Lebesgue measure 0) without changing the fact that $P_\theta[\delta^*(S) = 0] = 1$ for all $\theta$.]

12.3 Uniformly minimum variance estimators

In this section culminates the work of this chapter. First, a definition.

**Definition 20.** In the model with $X$, $X$, and parameter space $\Theta$, $\delta$ is a uniformly minimum variance unbiased estimator (UMVUE) of the function $g(\theta)$ if it is unbiased,

$$E_\theta[\delta(X)] = g(\theta) \text{ for all } \theta \in \Theta,$$  \hspace{1cm} (12.30)

has finite variance, and for any other unbiased estimator $\delta'$,

$$\text{Var}_\theta[\delta(X)] \leq \text{Var}_\theta[\delta'(X)] \text{ for all } \theta \in \Theta.$$ \hspace{1cm} (12.31)

And the theorem.

**Theorem 8.** Suppose $S = s(X)$ is a sufficient statistic for the model $(X, X, \Theta)$, and the model for $S$ is complete. Then if $\delta^*(s)$ is an unbiased estimator (depending on $s$) of the function $g(\theta)$, then $\delta_{\text{UMVUE}}(x) = \delta^*(s(x))$ is the UMVUE of $g(\theta)$.

**Proof.** Let $\delta$ be any unbiased estimator of $g(\theta)$, and consider

$$e_\delta(s) = E[\delta(X) \mid S = s].$$ \hspace{1cm} (12.32)

Because $S$ is sufficient, $e_\delta$ does not depend on $\theta$, so it is an estimator. Furthermore, since

$$E_\theta[e_\delta(S)] = E_\theta[\delta(X)] = g(\theta) \text{ for all } \theta \in \Theta,$$ \hspace{1cm} (12.33)

it is unbiased, and because it is a conditional expectation, as in (11.58),

$$\text{Var}_\theta[e_\delta(S)] \leq \text{Var}_\theta[\delta(X)] \text{ for all } \theta \in \Theta.$$ \hspace{1cm} (12.34)

But by completeness of the model for $S$, there is only one unbiased estimator that is a function of just $S$, $\delta^*$, hence $\delta^*(s) = e_\delta(s)$ (almost everywhere), and

$$\text{Var}_\theta[\delta^*(S)] \leq \text{Var}_\theta[\delta(X)] \text{ for all } \theta \in \Theta.$$ \hspace{1cm} (12.35)

This equation holds for any unbiased $\delta$, so that $\delta^*$ is best, or, as a function of $s$, $\delta_{\text{UMVUE}}(x) = \delta^*(s(x))$ is best. \hfill $\Box$

This proof is actually constructive in a sense. If you do not know $\delta^*$, but have any unbiased $\delta$, then you can find the UMVUE by using Rao-Blackwell’s (Theorem 7) conditioning on the sufficient statistic. Or, if you can by any means find an unbiased estimator that is a function of $S$, then it is UMVUE.
12.3.1 Example: Poisson
Consider again Example 12.2.1. We have that \( S = X_1 + \cdots + X_n \) is sufficient, and the model for \( S \) is complete. Because
\[
E_\theta[S/n] = \theta, \tag{12.36}
\]
we have that \( S/n \) is the UMVUE of \( \theta \).

Now let \( g(\theta) = e^{-\theta} = P_\theta[X_1 = 0] \). (If \( X_i \) is the number of phone calls in one minute, then \( g(\theta) \) is the chance there are 0 calls in the next minute.) We have from (12.13) an unbiased estimator of \( g(\theta) \) that is a function of \( S \), hence it is UMVUE. But finding the estimator took a bit of work, and luck. Instead, start with a very simple estimator,
\[
\delta(x) = I_{\{x_1 = 0\}}(x_1), \tag{12.37}
\]
which indicates whether there were 0 calls in the first minute. That estimator is unbiased, but obviously not using all the data. From (11.33) we have that \( X \) given \( S = s \) is multinomial, hence \( X_1 \) is binomial:
\[
X_1 \mid S = s \sim \text{Binomial}(s, \frac{1}{n}). \tag{12.38}
\]
Then
\[
E[\delta(X) \mid s(X) = s] = P[X_1 = 0 \mid S = s] = P[\text{Binomial}(s, \frac{1}{n}) = 0] = (1 - \frac{1}{n})^s. \tag{12.39}
\]
That is the UMVUE, and (as must be) it is the same as the estimator in (12.13).

12.4 Completeness for exponential families

Showing completeness of a model in general is not an easy task. Fortunately, exponential families as in (11.22) usually do provide completeness for the natural sufficient statistic. To present the result, we start assuming the model for the \( p \times 1 \) vector \( T \) has exponential family density with \( \theta \in \Theta \subset \mathbb{R}^p \) as the natural parameter, and \( T \) itself as the natural sufficient statistic:
\[
f_\theta(t) = a(t) e^{\theta_1 t_1 + \cdots + \theta_p t_p - \psi(\theta)}. \tag{12.40}
\]

Lemma 17. In the above model for \( T \), suppose that the parameter space contains a nonempty open \( p \)-dimensional rectangle. Then the model is complete.

If \( T \) is a lattice, e.g., the set of vectors containing all nonnegative integers, then the lemma can be proven by looking at the power series in \( e^{\theta_i} \)'s. In general, the proof would use uniqueness of Laplace transforms, as would the proof of the uniqueness of moment generating functions.

The requirement on the parameter space guards against exact constraints among the parameters. For example, suppose \( X_1, \ldots, X_n \) are iid \( N(\mu, \sigma^2) \), where \( \mu \) is known to be positive and the coefficient of variation, \( \sigma/\mu \), is known to be 10%. The exponent in the pdf, with \( \mu = 10\sigma \), is
\[
\frac{\mu}{\sigma^2} \sum x_i - \frac{1}{2\sigma^2} \sum x_i^2 = \frac{10}{\sigma} \sum x_i - \frac{1}{2\sigma^2} \sum x_i^2. \tag{12.41}
\]
The exponential family terms are then
\[ \theta = \left( \frac{10}{\sigma}, -\frac{1}{2\sigma^2} \right) \quad \text{and} \quad T = (\sum X_i, \sum X_i^2). \] (12.42)

The parameter space, with \( p = 2 \), is
\[ \Theta = \left\{ \left( \frac{10}{s}, -\frac{1}{2s^2} \right) \mid s > 0 \right\}, \] (12.43)

which does not contain a two-dimensional open rectangle, because \( \theta_2 = -\theta_1^2/100 \). Thus we cannot use the lemma to show completeness.

**NOTE.** It is important to read the lemma carefully. It does not say that if the requirement is violated, then the model is not complete. (Although realistic counterexamples are hard to come by.) To prove a model is not complete, you must produce a nontrivial unbiased estimator of \( \theta \). For example, in (12.42),
\[ E_{\theta_1}\left[ \left( T_1/n \right)^2 \right] = E_{\theta_1}[X^2] = \mu^2 + \frac{\sigma^2}{n} = (10n\sigma)^2 + \frac{\sigma^2}{n} = (100n^2 + \frac{1}{n})\sigma^2 \] (12.44)
and
\[ E_{\theta_1}[T_2/n] = E_{\theta_1}[X_i^2] = \mu^2 + \sigma^2 = (10n\sigma)^2 + \sigma^2 = (100n^2 + 1)\sigma^2. \] (12.45)

Then
\[ \delta(t_1, t_2) = \frac{t_1}{(100n^2 + \frac{1}{n})} - \frac{t_2}{(100n^2 + 1)} \] (12.46)

has expected value 0, but is not zero itself. Thus the model is not complete.

### 12.4.1 Examples

Suppose \( p = 1 \), so that the natural sufficient statistic and parameter are both scalars, \( T \) and \( \theta \), respectively. Then all that is needed is that \( T \) is not constant, and \( \Theta \) contains an interval \((a, b)\), \( a < b \). The table below has some examples, where \( X_1, \ldots, X_n \) are iid with the given distribution (the parameter space is assumed to be the most general):

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Sufficient statistic ( T )</th>
<th>Natural parameter ( \theta )</th>
<th>( \Theta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N(\mu, 1) )</td>
<td>( \sum X_i )</td>
<td>( \mu )</td>
<td>( \mathbb{R} )</td>
</tr>
<tr>
<td>( N(0, \sigma^2) )</td>
<td>( \sum X_i^2 )</td>
<td>( \frac{1}{2\sigma^2} )</td>
<td>( (0, \infty) )</td>
</tr>
<tr>
<td>Poisson(( \lambda ))</td>
<td>( \sum X_i )</td>
<td>( \log(\lambda) )</td>
<td>( \mathbb{R} )</td>
</tr>
<tr>
<td>Exponential(( \lambda ))</td>
<td>( \sum X_i )</td>
<td>( -\lambda )</td>
<td>( (-\infty, 0) )</td>
</tr>
</tbody>
</table>

The parameter spaces all contain open intervals; in fact they are open intervals. Thus the models for the \( T \)’s are all complete.

For \( X_1, \ldots, X_n \) iid \( N(\mu, \sigma^2) \), with \((\mu, \sigma^2) \in \mathbb{R} \times (0, \infty) \), the exponential family terms are
\[ \theta = \left( \frac{\mu}{\sigma}, -\frac{1}{2\sigma^2} \right) \quad \text{and} \quad T = (\sum X_i, \sum X_i^2). \] (12.48)

Here, without any extra constraints on \( \mu \) and \( \sigma^2 \), we have
\[ \Theta = \mathbb{R} \times (-\infty, 0), \] (12.49)
because for and \((a, b)\) with \(a \in \mathbb{R}, b < 0, \mu = a/\sqrt{1 - 2b} \in \mathbb{R}\) and \(\sigma^2 = -1/(2b) \in (0, \infty)\) are valid parameter values. Thus the model is complete for \((T_1, T_2)\). From the UMVUE Theorem 8, we then have that any function of \((T_1, T_2)\) is the UMVUE for its expected value. For example, \(\overline{X}\) is the UMVUE for \(\mu, S^2 = \sum(X_i - \overline{X})^2/(n - 1)\) is the UMVUE for \(\sigma^2\). Also, since \(E[(\overline{X})^2] = \mu^2 + \sigma^2/n\), the UMVUE for \(\mu^2\) is \((\overline{X})^2 - S^2/n\).

12.5 Fisher’s Information and the Cramér-Rao Lower Bound

If you find a UMVUE, then you know you have the best unbiased estimator. Often-times, there is no UMVUE, or there is one but it is too difficult to calculate. Then what? At least it would be informative to have an idea of whether a given estimator is very poor, or just not quite optimal. One approach is to find a lower bound for the variance. The closer an estimator is to the lower bound, the better.

The model for this section has \(X_1, X_2\) densities \(f_\theta(x)\), and parameter space \(\Theta = (a, b) \subset \mathbb{R}\). We need various assumptions, which we will postpone detailing until Section 17.1. Mainly, we need that the pdf is always positive (so that \(Unif(0, \theta)'s\) are not allowed), and has several derivatives with respect to \(\theta\), and that certain expected values are finite. Suppose \(\delta\) is an unbiased estimator of \(g(\theta)\), so that

\[
E_\theta[\delta(X)] = \int_X \delta(x)f_\theta(x)dx = g(\theta) \quad \text{for all } \theta \in \Theta. \tag{12.50}
\]

(Use a summation if the density is a pmf.) Now take the derivative with respect to \(\theta\) of both sides. Assuming the following steps are valid, we have

\[
g'(\theta) = \frac{\partial}{\partial \theta} \int_X \delta(x)f_\theta(x)dx
= \int_X \delta(x) \frac{\partial}{\partial \theta} f_\theta(x)dx
= \int_X \delta(x) \frac{\partial}{\partial \theta} \log(f_\theta(x)) f_\theta(x)dx
= E_\theta[\delta(X) \frac{\partial}{\partial \theta} \log(f_\theta(X))]. \tag{12.51}
\]

(Don’t ask how anyone decided to take these steps. That’s why these people are famous.)

The final term looks almost like a covariance. It is, because if we put \(\delta(x) = 1\) in the steps (12.51), since \(\int_X \delta(x)f_\theta(x)dx = 1\), and the derivative of 1 is 0,

\[
0 = E_\theta[\frac{\partial}{\partial \theta} \log(f_\theta(X))]. \tag{12.52}
\]

Then (12.51) and (12.52) show that

\[
g'(\theta) = \text{Cov}_\theta[\delta(X), \frac{\partial}{\partial \theta} \log(f_\theta(X))]. \tag{12.53}
\]

Now we can use the correlation inequality, \(\text{Cov}[U, V]^2 \leq \text{Var}[U] \text{Var}[V]\):

\[
g'(\theta)^2 \leq \text{Var}_\theta[\delta(X)] \text{Var}_\theta[\frac{\partial}{\partial \theta} \log(f_\theta(X))]. \tag{12.54}
\]
or
\[ \text{Var}_\theta[\delta(X)] \geq \frac{g'(\theta)^2}{\text{Var}_\theta[\frac{\partial}{\partial \theta} \log(f_\theta(X))]} . \] (12.55)

This equation is the key. The Cramér-Rao lower bound (CRLB) for unbiased estimators of \( g(\theta) \) is
\[ \text{CRLB}_g(\theta) = \frac{g'(\theta)^2}{I(\theta)} , \] (12.56)
where \( I \) is Fisher's Information in the model,
\[ I(\theta) = \text{Var}_\theta[\frac{\partial}{\partial \theta} \log(f_\theta(X))] . \] (12.57)

Under appropriate conditions, among them that \( I(\theta) \) is positive and finite, for any unbiased estimator \( \delta \) of \( g(\theta) \),
\[ \text{Var}_\theta[\delta(X)] \geq \frac{g'(\theta)^2}{I(\theta)} . \] (12.58)

Chapter 17 studies Fisher’s information in more depth. Notice that \( I(\theta) \) is a function just of the model, not of any particular estimator or estimand\(^1\). The larger it is, the better, because that means the variance of the estimator is potentially lower. So information is good.

If an unbiased estimator achieves the CRLB, that is,
\[ \text{Var}_\theta[\delta(X)] = \frac{g'(\theta)^2}{I(\theta)} \text{ for all } \theta \in \Theta , \] (12.59)
then it is the UMVUE. The converse is not necessarily true. The UMVUE may not achieve the CRLB, which of course means that no unbiased estimator does. There are other more accurate bounds, called Bhattacharya bounds, that the UMVUE may achieve in such cases.

### 12.5.1 Example: Double exponential

Suppose \( X_1, \ldots, X_n \) are iid Double Exponential(\( \theta \)), \( \theta \in \mathbb{R} \), so that
\[ f_\theta(x) = \frac{1}{2n} e^{-\Sigma |x_i - \theta|} . \] (12.60)

This density is not an exponential family one, and if \( n > 1 \), the sufficient statistic is the vector of order statistics, which does not have a complete model. Thus the UMVUE approach does not bear fruit. Because \( E_\theta[X_i] = \theta \), \( \bar{X} \) is an unbiased estimator of \( \theta \), with
\[ \text{Var}_\theta[\bar{X}] = \frac{\text{Var}_\theta[X_i]}{n} = \frac{2}{n} . \] (12.61)

Is this variance reasonable? We will compare it to the CRLB:
\[ \frac{\partial}{\partial \theta} \log(f_\theta(x)) = \frac{\partial}{\partial \theta} - \sum |x_i - \theta| = \sum \text{Sign}(x_i - \theta) , \] (12.62)

\(^1\)That which is being estimated.
because the derivative of $|x|$ is $+1$ if $x > 0$ and $-1$ if $x < 0$. (We are ignoring the possibility that $x_i = \theta$, where the derivative does not exist.) By symmetry of the density around $\theta$, each $\text{Sign}(X_i - \theta)$ has a probability of $\frac{1}{2}$ to be either $-1$ or $+1$, so it has mean 0 and variance 1. Thus

$$I(\theta) = \text{Var}_\theta[\sum \text{Sign}(X_i - \theta)] = n. \tag{12.63}$$

Here, $g(\theta) = \theta$, hence

$$\text{CRLB}_g(\theta) = \frac{g'(\theta)^2}{I(\theta)} = \frac{1}{n}. \tag{12.64}$$

Compare this bound to the variance in (12.61). The variance of $\bar{X}$ is twice the CRLB, which is not very good. It appears that there should be a better estimator. In fact, there is, which we will see next in Chapter 13, although even that estimator does not achieve the CRLB.

### 12.5.2 Example: Normal $\mu^2$

Suppose $X_1, \ldots, X_n$ are iid $N(\mu, 1)$, $\mu \in \mathbb{R}$. We know that $\bar{X}$ is sufficient, and the model for it is complete, hence any unbiased estimator based on the sample mean is UMVUE. In this case, $g(\mu) = \mu^2$, and $E_\mu[\bar{X}^2] = \mu^2 + 1/n$, hence

$$\delta(x) = x^2 - \frac{1}{n} \tag{12.65}$$

is the UMVUE. To find the variance of the estimator, start by noting that $\sqrt{n\bar{X}} \sim N(\sqrt{n\mu}, 1)$, so that its square is noncentral $\chi^2$ (Definition 10),

$$n\bar{X}^2 \sim \chi^2_1(n\mu^2). \tag{12.66}$$

Thus from (8.67),

$$\text{Var}_\mu[n\bar{X}^2] = 2 + 4n\mu^2 = 2\nu + 4\Delta. \tag{12.67}$$

Finally,

$$\text{Var}_\mu[\delta(X)] = \text{Var}_\mu[\bar{X}^2] = \frac{2}{n^2} + \frac{4\mu^2}{n}. \tag{12.68}$$

For the CRLB, we first need Fisher’s Information. Start with

$$\frac{\partial}{\partial \mu} \log(f_\mu(x)) = \frac{\partial}{\partial \mu} - \frac{1}{2} \sum (x_i - \mu)^2 = \sum (x_i - \mu). \tag{12.69}$$

The $X_i$’s are independent with variance 1, so that $I(\mu) = n$. Thus with $g(\mu) = \mu^2$,

$$\text{CRLB}_g(\mu) = \frac{(2\mu)^2}{n} = \frac{4\mu^2}{n}. \tag{12.70}$$

Comparing (12.68) to (12.70), we see that the UMVUE does not achieve the CRLB, which implies that no unbiased estimator will. But note that the variance of the UMVUE is only off by $2/n^2$, so that for large $n$, it is close to achieving the CRLB. The ratio

$$\frac{\text{CRLB}_g(\mu)}{\text{Var}_\mu[\delta(X)]} = \frac{1}{1 + \frac{1}{2n\mu^2}} \tag{12.71}$$

measures how close to ideal the estimator is, the closer to 1, the better. Letting $n \to \infty$, we obtain a measure of asymptotic efficiency, which in this case is 1.
13.1 Introduction

Unbiasedness is one criterion to use for choosing estimators. It is generally a desirable property, and restricting to unbiased estimators eases the search for good estimators. On the other hand, exact unbiasedness is sometimes too restrictive, e.g., there may be no unbiased estimators of a particular quantity, or slightly biased estimators (such as MLE’s) are more pleasing in other ways.

Other useful criteria go under the names invariance and equivariance. These are related properties but not identical; even so, they are often conflated. As an example of an equivariant estimator, suppose $X_1, \ldots, X_n$ are the observed heights in inches of a sample of adults, and $\overline{X}$ is the sample mean, which turns out to be 67 inches. Now change the scale of these heights to centimeters, so that the data are now $Y_1 = 2.54X_1, \ldots, Y_n = 2.54X_n$. What is the sample mean of the $Y_i$’s?

The idea is that if one multiplies all the data by the same nonzero constant $c$, the sample mean is also multiplied by $c$. The sample mean is said to be equivariant under scale transformations. The sample median is also scale equivariant.

An affine transformation is one which changes both location and scale, e.g., the change from degrees centigrade to degrees Fahrenheit. If $X_1, \ldots, X_n$ are measurements in degrees centigrade, then

$$ Y_i = 32 + \frac{9}{5} X_i $$

is in degrees Fahrenheit. The sample mean (and median) is again equivariant under this transformation, that is, $\overline{Y} = 32 + (9/5)\overline{X}$. A shrinkage estimator here would not be equivariant, e.g., $0.95 \times \overline{X}$ is not equivariant: $0.95\overline{Y} = 0.95 \times 32 + 0.95(9/5)\overline{X} \neq 32 + (9/5)(0.95\overline{X})$. An argument for equivariance is that the estimator should not depend on whether you are doing the calculations in the US or in Europe (or any other non-US place).

The technicalities needed in using invariance considerations start with a group of transformations that act on the data. This action induces an action on the distribu-
tion of the data. For example, if the data are iid \(N(\mu, \sigma^2)\), and the observations are multiplied by \(c\), then the distribution becomes iid \(N(c\mu, c^2\sigma^2)\). For invariance considerations to make sense, the model has to be invariant under the group. Then we can talk about invariant and equivariant functions. All these notions are explained in Section 13.5. The present chapter treats the special case of shift invariance as applied to what are termed “location families.”

### 13.2 Location families

A location family of distributions is one for which the only parameter is the center; the shape of the density stays the same. We will deal with the one-dimensional iid case, where the data are \(X_1, \ldots, X_n\) iid, the space for \(\mathbf{X} = (X_1, \ldots, X_n)\) is \(\mathbb{R}^n\), and the density of \(X_i\) is \(f(x_i - \theta)\) for some given density \(f\). The parameter space is \(\mathbb{R}\). Thus the density of \(\mathbf{X}\) is

\[
f_{\theta}(\mathbf{x}) = \prod_{i=1}^{n} f(x_i - \theta).
\]

This \(\theta\) may or may not be the mean or median.

Examples include the \(X_i\)'s being \(N(\theta, 1)\), or Uniform(\(\theta, \theta + 1\)), or Double Exponential (\(\theta\)) as in (12.60). By contrast, Uniform(0, \(\theta\)) and Exponential (\(\theta\)) are not location families since the spread as well as the center is affected by the \(\theta\).

A location family model is shift-invariant, because if we add the same constant \(a\) to all the \(X_i\)'s, the resulting random vector has the same model as \(\mathbf{X}\). That is, suppose \(a \in \mathbb{R}\) is a fixed, and we look at the transformation

\[
\mathbf{X}^* = (X_1^*, \ldots, X_n^*) = (X_1 + a, \ldots, X_n + a).
\]

The inverse transformation uses \(X_i = X_i^* - a\), and the \(|Jacobian| = 1\), so the density of \(\mathbf{Y}\) is

\[
f_{\theta}^*(\mathbf{x}^*) = \prod_{i=1}^{n} f(x_i^* - a - \theta) = \prod_{i=1}^{n} f(x_i^* - \theta^*), \quad \text{where } \theta^* = \theta + a.
\]

Thus adding \(a\) to everything just shifts everything by \(a\). Note that the space of \(\mathbf{X}^*\) is the same as the space of \(\mathbf{X}\), \(\mathbb{R}^n\), and the space for the parameter \(\theta^*\) is the same as that for \(\theta\), \(\mathbb{R}\):

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Model*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample space</td>
<td>(\mathbf{X})</td>
<td>(\mathbf{X}^*)</td>
</tr>
<tr>
<td>Parameter space</td>
<td>(\mathbb{R}^n)</td>
<td>(\mathbb{R}^n)</td>
</tr>
</tbody>
</table>
| Density | \(\prod_{i=1}^{n} f(x_i - \theta)\) | \(\prod_{i=1}^{n} f(x_i^* - \theta^*)\)

Thus the two models are the same. Not that \(\mathbf{X}\) and \(\mathbf{X}^*\) are equal, but that the sets of distributions considered for them are the same. This is the sense in which the model is shift-invariant.

**Note.** If the model includes a prior on \(\theta\), then the two models are not the same, because the prior distribution on \(\theta^*\) would be different than that on \(\theta\).
13.3 Shift-equivariant estimators

Suppose \( \delta(x) \) is an estimator of \( \theta \) in a location family model. Because the two models in (13.6) are exactly the same, \( \delta(x^*) \) would be an estimator of \( \theta^* \). Adding \( a \) to the \( x_i \)'s adds \( a \) to the \( \theta \), hence a “reasonable” estimator would behave similarly. That is,

\[
\text{If } \delta(x) \text{ is an estimator of } \theta, \\
\text{and} \\
\delta(x^*) = \delta((x_1 + a, \ldots, x_n + a)) \text{ is an estimator of } \theta^* = \theta + a,
\]

then we should have

\[
\delta((x_1 + a, \ldots, x_n + a)) = \delta((x_1, \ldots, x_n)) + a.
\]

(13.8)

An estimator satisfying (13.8) is called shift-equivariant. Shifting the data by \( a \) shifts the estimator by \( a \). Examples of shift-equivariant estimators include \( \bar{x} \), the median, and \( (\text{Min} + \text{Max})/2 \).

A typical estimator may have its variance, bias, or MSE depending on the parameter. E.g., in the binomial case, the variance of the MLE is \( p(1-p)/n \). It turns out that in the location family case, a shift-equivariant estimator’s variance, bias, and (hence) MSE are all independent of \( \theta \). The key is that we can normalize \( X \) to have parameter \( 0 \):

\[
X_1, \ldots, X_N \text{ iid } f_\theta \implies Z_1, \ldots, Z_n \text{ iid } f_0,
\]

where

\[
Z_1 = X_1 - \theta, \ldots, Z_n = X_n - \theta.
\]

(13.9)

Letting \( a = -\theta \) in (13.8),

\[
\delta(X_1, \ldots, X_n) = \delta(X_1 - \theta, X_n - \theta) + \theta = \delta(Z_1, \ldots, Z_n) + \theta.
\]

(13.11)

Then

\[
E_\theta[\delta(X_1, \ldots, X_n)] = E_0[\delta(Z_1, \ldots, Z_n)] + \theta,
\]

(13.12)

hence

\[
\text{Bias}(\theta; \delta) = E_\theta[\delta(X_1, \ldots, X_n)] - \theta = E_0[\delta(Z_1, \ldots, Z_n)],
\]

(13.13)

\[
\text{Var}_\theta[\delta(X)] = E_\theta[(\delta(X) - E_\theta[\delta(X)])^2] = E_0[\delta(Z) - E_0[\delta(Z)]^2] = \text{Var}_0[\delta(Z)],
\]

(13.14)

and

\[
\text{MSE}(\theta; \delta) = \text{Var}_\theta[\delta(X)] + \text{Bias}^2(\theta; \delta) = \text{Var}_0[\delta(Z)] + E_0[\delta(Z)]^2 = E_0[\delta(Z)]^2.
\]

(13.15)

Thus when comparing shift-equivariant estimators, it is enough to look at the MSE when \( \theta = 0 \). The next subsection finds the best such estimator.

13.3.1 The Pitman estimator

By the best shift-equivariant estimator, we mean the equivariant estimator with the lowest mean square error for all \( \theta \). It is analogous to the UMVUE, the unbiased estimator with the lowest variance \( \equiv \text{MSE} \).

For any biased equivariant estimator, it is fairly easy to find one that is unbiased but has the same variance by just shifting it a bit. For example, suppose that \( X_i \)'s are iid \( \text{Uniform}(\theta, \theta + 1) \). Let \( \delta(x) = \bar{x} \). Then

\[
E_\mu[\delta(X)] = E_\theta[X_i] = \theta + \frac{1}{2} \quad \text{and} \quad \text{Var}_\theta[\delta(X)] = \frac{\text{Var}_0[X_i]}{n} = \frac{1}{12n},
\]

(13.16)
hence the bias is $E_0[\delta(\bar{X})] = \frac{1}{2}$, and

$$MSE_\theta[\delta(\bar{X})] = \frac{1}{12n} + \frac{1}{4}.$$  \hspace{1cm} (13.17)

But we can de-bias the estimator easily enough by subtracting the bias $\frac{1}{2}$, giving $\delta^*(\bar{X}) = \bar{X} - \frac{1}{2}$. Then the bias is 0, but the variance is the same as for $\delta$, hence,

$$MSE_\theta[\delta^*] = \frac{1}{12n} < \frac{1}{12n} + \frac{1}{4} = MSE_\theta[\delta].$$  \hspace{1cm} (13.18)

That works for any shift-equivariant estimator $\delta$, i.e., suppose $\delta$ is biased, and $b = E_0[\delta(Z)]$ is the bias. Then $\delta^* = \delta - b$ is also shift-equivariant, with

$$Var_0[\delta^*(Z)] = Var_0[\delta(Z)]$$ and $E_0[\delta^*(Z)] = 0,$ \hspace{1cm} (13.19)

hence

$$MSE(\theta; \delta^*) = Var_0[\delta(Z)] < Var_0[\delta(Z)] + E_0[\delta(Z)]^2 = MSE(\theta; \delta).$$  \hspace{1cm} (13.20)

Thus if a shift-equivariant estimator is biased, it can be improved, ergo ...

**Lemma 18.** If $\delta$ is the best shift-equivariant estimator, then it is unbiased.

Notice that this lemma implies that in the normal case, $\bar{X}$ is the best shift-equivariant estimator, because it is the best unbiased estimator, and it is shift-equivariant. But, of course, in the normal case, $\bar{X}$ is always the best at everything.

Now consider the general location family case. To find the best estimator, we first characterize the shift-equivariant estimators. We can use a similar trick as we did when finding the UMVUE when we took a simple unbiased estimator, then found its expected value given the sufficient statistic. Sufficiency is not very interesting, because for most $f$, there is no complete sufficient statistic. (The normal and uniform are a couple of exceptions.) But we start with an arbitrary shift-equivariant estimator $\delta$, and let “$X_n$” be another shift-equivariant estimator, and look at the difference:

$$\delta(x_1, \ldots, x_n) - x_n = \delta(x_1 - x_n, x_2 - x_n, \ldots, x_{n-1} - x_n, 0).$$  \hspace{1cm} (13.21)

Define the function $v$ on the differences $y_i = x_i - x_n$,

$$v : \mathbb{R}^{n-1} \longrightarrow \mathbb{R}$$

$$y \longrightarrow v(y) = \delta(y_1, \ldots, y_{n-1}).$$  \hspace{1cm} (13.22)

Then what we have is that for any equivariant $\delta$, there is a $v$ such that

$$\delta(\bar{X}) = x_n + v(x_1 - x_n, \ldots, x_{n-1} - x_n).$$  \hspace{1cm} (13.23)

Now instead of trying to find the best $\delta$, we will look for the best $v$, then use (13.23) to get the $\delta$. From (13.15), the best $v$ is found by minimizing

$$E_0[\delta(Z)^2] = E_0[(Z_n + v(Z_1 - Z_n, \ldots, Z_{n-1} - Z_n))^2].$$  \hspace{1cm} (13.24)

The trick is to condition on the differences $Z_1 - Z_n, \ldots, Z_{n-1} - Z_n$:

$$E_0[(Z_n + v(Z_1 - Z_n, \ldots, Z_{n-1} - Z_n))^2] = E_0[e(Z_1 - Z_n, \ldots, Z_{n-1} - Z_n)],$$  \hspace{1cm} (13.25)
where
\[
e(y_1, \ldots, y_{n-1}) = E_0[(Z_n + v(Z_1 - Z_n, \ldots, Z_{n-1} - Z_n))^2 | Z_1 - Z_n = y_1, \ldots, Z_{n-1} - Z_n = y_{n-1}]
\] (13.26)
\[
= E_0[(Z_n + v(y_1, \ldots, y_{n-1}))^2 | Z_1 - Z_n = y_1, \ldots, Z_{n-1} - Z_n = y_{n-1}].
\] (13.27)

It is now possible to minimize \( e \) for each fixed set of \( y_i \)’s, e.g., by differentiating with respect to \( v(y_1, \ldots, y_{n-1}) \). But we know the minimum is minus the (conditional) mean of \( Z_n \), that is, the best \( v \) is
\[
v(y_1, \ldots, y_{n-1}) = -E[Z_n | Z_1 - Z_n = y_1, \ldots, Z_{n-1} - Z_n = y_{n-1}].
\] (13.29)

That is a straightforward calculation. We just need the joint pdf divided by the marginal to get the conditional pdf, then find the expected value. For the joint, let
\[
Y_1 = Z_1 - Z_n, \ldots, Y_{n-1} = Z_{n-1} - Z_n,
\] (13.30)
and find the pdf of \((Y_1, \ldots, Y_{n-1}, Z_n)\). We use the Jacobian approach, so need the inverse function:
\[
z_1 = y_1 + z_n, \ldots, z_{n-1} = y_{n-1} + z_n, z_n = z.
\] (13.31)

Taking the \( \partial z_i / \partial y_i \)’s, etc., we have
\[
J = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 & 1 \\
0 & 1 & 0 & \cdots & 0 & 1 \\
0 & 0 & 1 & \cdots & 0 & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 1 \\
0 & 0 & 0 & \cdots & 0 & 1
\end{pmatrix}
\] (13.32)

The determinant of \( J \) is just 1, then, so the pdf of \((Y_1, \ldots, Y_{n-1}, Z_n)\) is
\[
f^* (y_1, \ldots, y_{n-1}, z_n) = f(z_n) \prod_{i=1}^{n-1} f(y_i + z_n),
\] (13.33)
and the marginal of the \( Y_i \)’s is
\[
f_Y^*(y_1, \ldots, y_{n-1}) = \int_{-\infty}^{\infty} f(z_n) \prod_{i=1}^{n-1} f(y_i + z_n)dz_n.
\] (13.34)

The conditional pdf:
\[
f^*(z_n | y_1, \ldots, y_{n-1}) = \frac{f(z_n) \prod_{i=1}^{n-1} f(y_i + z_n)}{\int_{-\infty}^{\infty} f(z_n) \prod_{i=1}^{n-1} f(y_i + z_n)dz_n}.
\] (13.35)

The conditional mean is then
\[
E[Z_n | Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1}] = \frac{\int_{-\infty}^{\infty} z_n f(z_n) \prod_{i=1}^{n-1} f(y_i + z_n)dz_n}{\int_{-\infty}^{\infty} f(z_n) \prod_{i=1}^{n-1} f(y_i + z_n)dz_n}
= -v(y_1, \ldots, y_{n-1}).
\] (13.36)
The best δ uses that ν, so that from (13.23), with \( y_i = x_i - x_n \),

\[
\delta(x) = x_n - \frac{\int_{-\infty}^{\infty} z_n f(z_n) \prod_{i=1}^{n-1} f(x_i - x_n + z_n) dz_n}{\int_{-\infty}^{\infty} f(z_n) \prod_{i=1}^{n-1} f(x_i - x_n + z_n) dz_n}
\]

\[
= \frac{\int_{-\infty}^{\infty} (x_n - z_n) f(z_n) \prod_{i=1}^{n-1} f(x_i - x_n + z_n) dz_n}{\int_{-\infty}^{\infty} f(z_n) \prod_{i=1}^{n-1} f(x_i - x_n + z_n) dz_n}.
\]

(13.37)

Just to make it look better, make the change-of-variables \( \theta = x_n - z_n \), so that \( z_n = x_n - \theta \), and

\[
\delta(x) = \frac{\int_{-\infty}^{\infty} \theta f(x_n - \theta) \prod_{i=1}^{n-1} f(x_i - \theta) d\theta}{\int_{-\infty}^{\infty} f(x_n - \theta) \prod_{i=1}^{n-1} f(x_i - \theta) d\theta}
\]

\[
= \frac{\int_{-\infty}^{\infty} \theta \prod_{i=1}^{n-1} f(x_i - \theta) d\theta}{\int_{-\infty}^{\infty} \prod_{i=1}^{n-1} f(x_i - \theta) d\theta}.
\]

(13.38)

That’s the best, called the Pitman Estimator:

**Theorem 9.** In the location-family model with individual density \( f \), if there is an equivariant estimator with finite MSE, then it is given by the Pitman Estimator (13.38).

The final expression in (13.38) is very close to a Bayes posterior mean. With prior \( \pi \), the posterior mean is

\[
E[\theta | X = x] = \frac{\int_{-\infty}^{\infty} \theta f_\theta(x) \pi(\theta) d\theta}{\int_{-\infty}^{\infty} f_\theta(x) \pi(\theta) d\theta}.
\]

(13.39)

Thus the Pitman Estimator can be thought of as the posterior mean for prior \( \pi(\theta) = 1 \). Unfortunately, that \( \pi \) is not a pdf, because it integrates to \(+\infty\). Such a prior is called an improper prior.

### 13.4 Examples

One thing special about this theorem is that it is constructive, that is, there is a formula telling exactly how to find the best. By contrast, when finding the UMVUE, you have to do some guessing. If you find an unbiased estimator that is a function of a complete sufficient statistic, then you have it, but it may not be easy to find. You might be able to use the power-series approach, or find an easy one then use Rao-Blackwell, but maybe not. A drawback is that the integrals may not be easy to perform analytically, although in practice it would be straightforward to use numerical integration. The homework has a couple of example, the Normal and uniform, in which the integrals are doable. Here are additional examples:

#### 13.4.1 Exponential

Here, \( f(x) = e^{-x}I[x > 0] \), i.e., the Exponential(1) pdf. The location family is not Exponential(\( \theta \)), but a shifted Exponential(1) that starts at \( \theta \) rather than at 0:

\[
f(x - \theta) = e^{-(x-\theta)}I[x - \theta > 0] = \begin{cases} e^{-(x-\theta)} & \text{if} & x > \theta \\ 0 & \text{otherwise} \end{cases}.
\]

(13.40)
13.4. Examples

Then the pdf of $X$ is

$$f(x | \theta) = \prod_{i=1}^{n} e^{-(x_i - \theta)} I[x_i > \theta] = e^{n\theta} e^{-\Sigma x_i} I[\min \{x_i\} > \theta], \quad (13.41)$$

because all $x_i$’s are greater than $\theta$ if and only if the minimum is. Then the Pitman estimator is, from (13.38),

$$\delta(x) = \frac{\int_{-\infty}^{\infty} \theta e^{n\theta} e^{-\Sigma x_i} I[\min \{x_i\} > \theta] d\theta}{\int_{-\infty}^{\infty} e^{n\theta} e^{-\Sigma x_i} I[\min \{x_i\} > \theta] d\theta} = \frac{\int_{-\infty}^{m} \theta e^{n\theta} d\theta}{\int_{-\infty}^{m} e^{n\theta} d\theta}, \quad (13.42)$$

where $m = \min \{x_i\}$. Do the numerator using integration by parts, so that

$$\int_{-\infty}^{m} \theta e^{n\theta} d\theta = \frac{\theta}{n} e^{n\theta} |_{-\infty}^{m} - \frac{1}{n} \int_{-\infty}^{m} e^{n\theta} d\theta = \frac{m}{n} e^{mn} - \frac{1}{n^2} e^{mn}, \quad (13.43)$$

and the denominator is $\left(1/n\right)e^{mn}$, hence

$$\delta(x) = \frac{\frac{m}{n} e^{mn} - \frac{1}{n^2} e^{mn}}{\frac{1}{n} e^{mn}} = \min \{x_i\} - \frac{1}{n}. \quad (13.44)$$

Is that estimator unbiased? Yes, because it is the best equivariant estimator. Actually, it is the UMVUE, because $\min \{x_i\}$ is a complete sufficient statistic.

13.4.2 Double Exponential

Now

$$f(x | \theta) = \prod_{i=1}^{n} \frac{1}{2} e^{-|x_i - \theta|} = \frac{1}{2^n} e^{-\Sigma |x_i - \theta|}, \quad (13.45)$$

and the Pitman Estimator is

$$\delta(x) = \frac{\int_{-\infty}^{\infty} \theta e^{-\Sigma |x_i - \theta|} d\theta}{\int_{-\infty}^{\infty} e^{-\Sigma |x_i - \theta|} d\theta}. \quad (13.46)$$

These integrals need to be broken up, depending on which $(x_i - \theta)$’s are positive and which negative. We’ll do the $n = 2$ case in detail. First note that we get the same result if we use the order statistics, that is,

$$\delta(x) = \frac{\int_{-\infty}^{\infty} \theta e^{-\Sigma |x_{(i)} - \theta|} d\theta}{\int_{-\infty}^{\infty} e^{-\Sigma |x_{(i)} - \theta|} d\theta}. \quad (13.47)$$

So with $n = 2$, we have three regions of integration:

$$\begin{align*}
\theta < x_{(1)} & \Rightarrow \Sigma |x_{(i)} - \theta| = x_{(1)} + x_{(2)} - 2\theta; \\
x_{(1)} < \theta < x_{(2)} & \Rightarrow \Sigma |x_{(i)} - \theta| = -x_{(1)} + x_{(2)}; \\
x_{(2)} < \theta & \Rightarrow \Sigma |x_{(i)} - \theta| = -x_{(1)} - x_{(2)} + 2\theta. \quad (13.48)
\end{align*}$$
The numerator in (13.47) is
\[
e^{-x(1)-x(2)} \int_{-\infty}^{\infty} \theta e^{2\theta} d\theta + e^{x(1)-x(2)} \int_{x(1)}^{\infty} \theta d\theta + e^{x(1)+x(2)} \int_{x(2)}^{\infty} e^{-2\theta} d\theta
\]
\[
= e^{-x(1)-x(2)} (\frac{x(1)}{2} - \frac{1}{4}) e^{2x(1)} + e^{x(1)-x(2)} \frac{x(2)^2 - x(1)^2}{2} + e^{x(1)+x(2)} (\frac{x(2)}{2} + \frac{1}{4}) e^{-2x(2)}
\]
\[
= \frac{1}{2} e^{x(1)-x(2)} (x(1) + x(2)^2 - x(1)^2 + x(2)).
\]

The denominator is
\[
e^{-x(1)-x(2)} \int_{-\infty}^{\infty} e^{2\theta} d\theta + e^{x(1)-x(2)} \int_{x(1)}^{\infty} d\theta + e^{x(1)+x(2)} \int_{x(2)}^{\infty} e^{-2\theta} d\theta
\]
\[
= e^{-x(1)-x(2)} \frac{1}{2} e^{2x(1)} + e^{x(1)-x(2)} (x(2) - x(1)) + e^{x(1)+x(2)} \frac{1}{2} e^{-2x(2)}
\]
\[
= e^{x(1)-x(2)} (1 + x(2) - x(1)).
\]

Finally,
\[
\delta(x) = \frac{1}{2} \frac{x(1) + x(2)^2 - x(1)^2 + x(2)}{1 + x(2) - x(1)} = \frac{x(1) + x(2)}{2},
\]
which is a rather long way to calculate the mean! The answer for \( n = 3 \) is not the mean. Is it the median?

### 13.5 General invariance and equivariance

This section sketches the general notions of invariance and equivariance, expanding on the shift-invariance in the previous sections.

#### 13.5.1 Groups and their actions

See Section 13.5.5 for a brief review of groups. We are interest in how members of a group act on the data \( \mathbf{X} = (X_1, \ldots, X_n)' \) with space \( \mathcal{X} \). That is, given a group \( \mathcal{G} \), its action on \( \mathcal{X} \) is expressed by associating for each \( g \in \mathcal{G} \) a one-to-one and onto function \( h_g \),
\[
h_g : \mathcal{X} \longrightarrow \mathcal{X},
\]
so that the action of \( g \) on \( \mathbf{x} \) is the transformation \( h_g(\mathbf{x}) \). Some examples:

**Scale** The group is \( \mathcal{G} = \{c \in \mathbb{R} \mid c \neq 0\} \) under multiplication. It acts on the data by multiplying the observations by an element of the group, that is, for \( c \in \mathcal{G} \), the transformation is given by
\[
h_c : \mathbb{R}^n \longrightarrow \mathbb{R}^n
\]
\[
\mathbf{x} \longrightarrow h_c(\mathbf{x}) = (cx_1, \ldots, cx_n)' = c \mathbf{x}.
\]

**Shift** The group is \( \mathcal{G} = \mathbb{R} \) under addition. It acts on the data by adding a constant to each of the observations:
\[
h_a : \mathbb{R}^n \longrightarrow \mathbb{R}^n
\]
\[
\mathbf{x} \longrightarrow h_a(\mathbf{x}) = (a + x_1, \ldots, a + x_n)' = a\mathbf{1}_n + \mathbf{x},
\]
where \( \mathbf{1}_n = (1, \ldots, 1)' \), the column vector of all 1’s.
Affine The group is $G = \{(a, b) \in \mathbb{R}^2 \mid b \neq 0\}$. It acts on the data by multiplying the observations by $b$ and adding $a$:

$$h_{(a,b)} : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$\mathbf{x} \rightarrow h_{(a,b)}(\mathbf{x}) = (a + b \mathbf{x}_1, \ldots, a + b \mathbf{x}_n)' = a \mathbf{1}_n + b \mathbf{x}. \quad (13.55)$$

The temperature transformation in (13.2) is an example of an affine transformation.

The operation for the group is, for $(a, b)$ and $(c, d) \in G$,

$$(a, b) \circ (c, d) = (a + bc, bd), \quad (13.56)$$

so that

$$h_{(a,b) \circ (c,d)}(\mathbf{x}) = h_{(a,b)}(h_{(c,d)}(\mathbf{x}))$$

$$= h_{(a,b)}(c \mathbf{1}_n + d \mathbf{x})$$

$$= a \mathbf{1}_n + b(c \mathbf{1}_n + d \mathbf{x})$$

$$= (a + bc) \mathbf{1}_n + (bd) \mathbf{x}$$

$$= h_{(a+bc,bd)}(\mathbf{x}) \quad (13.57)$$

Reflection The group is $G = \{-1, +1\}$ under multiplication. It acts on the data by multiplication:

$$h_\varepsilon : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$\mathbf{x} \rightarrow h_\varepsilon(\mathbf{x}) = \varepsilon \mathbf{x}. \quad (13.58)$$

Thus, it either leaves the data alone, or it switches all the signs.

Permutation The group is $G = \{P \mid P$ is an $n \times n$ permutation matrix$\}$ under multiplication. It acts on the data by multiplication:

$$h_P : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$\mathbf{x} \rightarrow h_P(\mathbf{x}) = P \mathbf{x}. \quad (13.59)$$

A permutation matrix has exactly one 1 in each row and one 1 in each column, and the rest of the elements are 0. Then $P \mathbf{x}$ just permutes the elements of $\mathbf{x}$. For example,

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad (13.60)$$

is a $4 \times 4$ permutation matrix, and

$$P \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_4 \\ x_3 \\ x_1 \end{pmatrix}. \quad (13.61)$$
The group actions need to cohere in a proper way. That is, in order for the $h_g$’s to be an action, they must satisfy the following:

- $h_ι(x) = x$, where $ι$ is the identity.
- $h_{g1 \circ g2}(x) = h_{g1}(h_{g2}(x))$ for any $g1, g2 \in G$.

In the above examples, it is fairly easy to see that the transformations do satisfy these requirements. For example, in the shift case, the identity is 0, and adding 0 to $x$ leaves it alone. Also, adding $a$ to the $x_i$’s, then adding $b$ to the $x_i + a$’s, is the same as adding $a + b$ to the $x_i$’s. The only example that is a little tricky in the affine one, and (13.57) shows that the second requirement works.

Whether a given group can be used for a given model depends on whether the transformed data remain in the model. This question is addressed next.

### 13.5.2 Invariant models

Start with the model consisting of random vector $X$, space $\mathcal{X}$, and set of distributions $\mathcal{P}$.

**Definition 21.** Suppose $G$ acts on $\mathcal{X}$, and that for any $g \in G$, if the distribution of $X$ is in $\mathcal{P}$, then the distribution of $h_g(X)$ is in $\mathcal{P}$. Then the model $(\mathcal{X}, \mathcal{P})$ is said to be invariant under the group $G$.

The definition is not saying that $X$ and $h_g(X)$ have the same distribution, but that they have the same type of distribution. Now for some examples.

1. Suppose $X \sim N(\mu, 1)$ where $\mu \in \mathbb{R}$, so the mean can be anything, but the variance has to be 1. First take $G_1 = \mathbb{R}$ under addition, so the action is the shift (13.54), with $n = 1$. Now $X^* = h_0(X) = a + X \sim N(a + \mu, 1)$, which is $N(\mu^*, 1)$ with $\mu^* = a + \mu \in \mathbb{R}$. So the model is invariant under $G_1$.

Next, consider the scale group, $G_2 = \{c \in \mathbb{R} \mid c \neq 0\}$ under multiplication. Then $X^* = g_c(X) = cX \sim N(c\mu, c^2)$. This $X^*$ does not have a distribution in the model for all $c$, since that model requires the variance to be 1.

But with the reflection group $G_3 = \{-1, +1\}$, $eX \sim N(e\mu, e^2) = N(e\mu, 1)$ for $e = \pm 1$, which is in the model. So the model is invariant under $G_3$.

2. Now let $X \sim N(\mu, \sigma^2)$ where $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$, so the mean and variance can be anything. This model is invariant under all the groups in the previous example. In fact, it is invariant under the affine group, since if $(a, b) \in G = \{(a, b) \in \mathbb{R}^2 \mid b \neq 0\}$, then $X^* = a + bX \sim N(a + b\mu, b^2\sigma^2) = N(\mu^*, \sigma^{*2})$, and $(\mu^*, \sigma^{*2}) \in \mathbb{R} \times (0, \infty)$ as required.

3. Suppose $X_1, \ldots, X_n$ are iid with any location-family distribution, that is, for a given pdf $f$, the pdf of $X$ is

$$\prod_{i=1}^{n} f(x_i - \mu),$$  \hspace{1cm} (13.62)

where $\mu \in \mathbb{R}$. This model is shift-invariant, since for $a \in \mathbb{R}$, $X^* = (a + X_1, \ldots, a + X_n)'$ has pdf

$$\prod_{i=1}^{n} f(x_i^* - \mu^*),$$  \hspace{1cm} (13.63)
where $\mu^* = a + \mu \in \mathbb{R}$.

4. Suppose $X_1, \ldots, X_n$ are iid with any distribution individually. Then the model is invariant under the permutation group, because any permutation of the $X_i$'s also has iid elements.

5. On the other hand, consider a little ANOVA model, where $X_1, X_2$ are $N(\mu, \sigma^2)$, and $X_3, X_4$ are $N(\nu, \sigma^2)$, and the $X_i$'s are independent. The parameter space is $(\mu, \nu) \in \mathbb{R}^2$. So

$$X \sim N\left(\begin{pmatrix} \mu \\ \mu \\ \nu \\ \nu \end{pmatrix}, \sigma^2 I_4\right).$$

This model is not invariant under the permutation group. E.g., take $P$ as in (13.60), so that from (13.61),

$$PX = \begin{pmatrix} X_2 \\ X_4 \\ X_3 \\ X_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \nu \\ \nu \\ \mu \end{pmatrix}, \sigma^2 I_4\right),$$

This mean does not have the right pattern (first two equal, last two equal), hence the distribution of $PX$ is not in the model. (There are permutations which do keep the distribution in the model, but to be invariant, it has to work for every permutation.)

13.5.3 Induced actions

Most of the models we’ll consider are parametric, i.e., $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$. If a parametric model is invariant under a group $G$, then for $g \in G$,

$$X \sim P_\theta \implies X^* = h_g(X) \sim P_{\theta^*},$$

for some other parameter $\theta^* \in \Theta$. In most of the examples above, that is what happened. For example, #2. where $X \sim N(\mu, \sigma^2)$, $(\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. The group is the affine group. Then for $(a, b) \in G$, $X^* = a + bX$, and $(\mu^*, \sigma^{*2}) = (a + b\mu, b^2 \sigma^2)$. That is, the group also acts on $\Theta$. It is a different action than that on $X$. Call the transformations $\bar{h}_g$, so that

$$\bar{h}_{(a,b)} : \Theta \rightarrow \Theta, \quad (\mu, \sigma^2) \rightarrow \bar{h}_{(a,b)}(\mu, \sigma^2) = (a + b\mu, b^2 \sigma^2).$$

This action is called the induced action on $\Theta$, because it automatically arises once the action on $X$ is defined.

It may be that the induced action is trivial, that is, the parameter does not change. For example, take $X \sim N(\mu 1_n, I_n)$, $\mu \in \mathbb{R}$. This model is invariant under permutations. But for a permutation matrix $P$, $PX \sim N(P(\mu 1_n), P I_n P') = N(\mu 1_n, I_n)$, so $\mu^* = \mu$, i.e., $\bar{h}_P(\mu) = \mu$. 

**13.5.4 Equivariant estimators**

The first application of invariance will be to estimating the location parameter in location families. The data are as in Example 3 in Section 13.5.2, so that \( X = (X_1, \ldots, X_n)' \), where the \( X_i \)'s are iid \( f(x_i - \mu) \), yielding the joint pdf (13.62). The parameter space is \( \mathbb{R} \). We wish to find the best shift-equivariant estimator of \( \mu \), but first we need to define what that means. Basically, it means the induced action on the estimator is the same as that on the parameter. The following definition is for the general case.

**Definition 22.** Suppose that the model with space \( \mathcal{X} \) and parameterized set of distributions \( \{ P_\theta | \theta \in \Theta \} \) is invariant under the group \( G \). Then an estimator \( \delta \) of \( \theta \) with

\[
\delta : \mathcal{X} \longrightarrow \Theta
\]

is equivariant under \( G \) if

\[
\delta(h \circ g(x)) = \tilde{h} \circ \delta(x)
\]

for each \( g \in G \).

**13.5.5 Groups**

**Definition 23.** A group \( G \) is a set of elements equipped with a binary operation \( \circ \) such that

- **Closure** If \( g_1, g_2 \in G \) then \( g_1 \circ g_2 \in G \).
- **Associativity** \( g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3 \) for any \( g_1, g_2, g_3 \in G \).
- **Identity** There exists an element \( \iota \in G \), called the identity, such that for every \( g \in G \), \( g \circ \iota = \iota \circ g = g \).
- **Inverse** For each \( g \in G \) there exists an element \( g^{-1} \in G \), called the inverse of \( g \), such that \( g \circ g^{-1} = g^{-1} \circ g = \iota \).

Some examples:

<table>
<thead>
<tr>
<th>Name</th>
<th>( G )</th>
<th>Element</th>
<th>( \circ )</th>
<th>Identity</th>
<th>Inverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shift</td>
<td>( \mathbb{R} )</td>
<td>( a )</td>
<td>+</td>
<td>0</td>
<td>(-a)</td>
</tr>
<tr>
<td>Scale</td>
<td>( \mathbb{R} - {0} )</td>
<td>( b )</td>
<td>\times</td>
<td>1</td>
<td>( \frac{1}{b} )</td>
</tr>
<tr>
<td>Affine</td>
<td>( \mathbb{R} \times [\mathbb{R} - {0}] )</td>
<td>( (a, b) )</td>
<td>( (a, b) \circ (c, d) = (a + bc, bd) )</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Reflection</td>
<td>( {-1, +1} )</td>
<td>( e )</td>
<td>\times</td>
<td>1</td>
<td>( e )</td>
</tr>
<tr>
<td>Permutation</td>
<td>( n \times n ) permutation matrices</td>
<td>( P )</td>
<td>\times</td>
<td>( I_n )</td>
<td>( P' )</td>
</tr>
</tbody>
</table>

\(* = \text{Exercise for the reader.}\)


Chapter 14

The Decision-theoretic Approach

14.1 The setup

Assume we have a statistical model, with $X$, $X'$, and $\Theta$. The decision-theoretic approach to inference supposes an action space $A$ that specifies the possible “actions” we might take. Some typical examples:

<table>
<thead>
<tr>
<th>Inference</th>
<th>Action space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimating $\theta$</td>
<td>$A = \Theta$</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>$A = {\text{Accept } H_0, \text{ Reject } H_0}$ (14.1)</td>
</tr>
<tr>
<td>Selecting among several models</td>
<td>$A = {\text{Model I, Model II, } \ldots}$</td>
</tr>
<tr>
<td>Predicting $X_{n+1}$</td>
<td>$A =$ Space of $X_n$</td>
</tr>
</tbody>
</table>

A decision procedure specifies which action to take for each possible value of the data. In estimation, the decision procedure is the estimator; in hypothesis testing, it is the hypothesis test. Formally, a decision procedure is a function $\delta(x)$,

$$\delta : \mathcal{X} \longrightarrow A.$$  \hfill (14.2)

[The above is a nonrandomized decision procedure. A randomized procedure would depend on not just the data $x$, but also some outside randomization element. We will ignore randomized procedures until Chapter 18 on hypothesis testing.]

A good procedure is one that takes good actions. To measure how good, we need a loss function that specifies a penalty for taking a particular action when a particular distribution obtains. Formally, the loss function $L$ is a function

$$L : A \times \Theta \longrightarrow [0, \infty).$$  \hfill (14.3)

When estimating a function $g(\theta)$, common loss functions are squared-error loss,

$$L(a, \theta) = (a - g(\theta))^2,$$  \hfill (14.4)

and absolute-error loss,

$$L(a, \theta) = |a - g(\theta)|.$$  \hfill (14.5)

In hypothesis testing, a “0 – 1” loss is common, where you lose 0 by making the correct decision, and lose 1 if you make a mistake.
14.2 The risk function

A frequentist evaluates procedures by their behavior over experiments. In this decision-theoretic framework, the risk function for a particular decision procedure $\delta$ is key. The risk is the expected loss, where $\delta$ takes place of the $a$. It is a function of $\theta$:

$$R(\theta; \delta) = E[L(\delta(X), g(\theta)) | \theta = \theta].$$  \hfill (14.6)

The two expectations are the same, but the first is written for frequentists, and the second for Bayesians.

In this chapter, we will concentrate on estimation, where $L$ is square-error loss (14.4), which means the risk is the mean square error from Definition 14:

$$R(\theta; \delta) = E[\delta(X) - g(\theta)^2] = MSE(\theta; \delta).$$  \hfill (14.7)

The idea is to choose a $\delta$ with small risk. The challenge is that usually, there is not a procedure that is best for every $\theta$. For example, in Section 10.6, we exhibit several estimators of the binomial parameter. Notice that the risk functions cross, so there is no way to say one procedure is “best.” One way to choose is to restrict consideration to a subset of estimators, e.g., unbiased ones as in Chapters 12, or shift-equivariant ones as in Chapter 13. Often, one either does not wish to use such restriction, or cannot.

In the absence of a uniquely defined best procedure, frequentists have several possible strategies, among them determining admissible procedures, minimax procedures, or Bayes procedures.

14.3 Admissibility

Recall the example in Section 10.6, where $X \sim Binomial(n, \theta), n = 5$. Figure 10.2, repeated here in Figure 14.1, shows $MSE(\theta; \delta)$ for four estimators.

The important message in the graph is that none of the four estimators is obviously “best” in terms of MSE. In addition, none of them are discardable in the sense that another estimator is better. That is, for any pair of estimators, sometimes one is better, sometimes the other. A fairly weak criterion is this lack of discardability. Formally, the estimator $\delta^*$ is said to dominate the estimator $\delta$ if

$$R(\theta; \delta^*) \leq R(\theta; \delta) \quad \text{for all } \theta \in \Theta,$$

$$R(\theta; \delta^*) < R(\theta; \delta) \quad \text{for at least one } \theta \in \Theta.$$ \hfill (14.8)

For example, suppose $X$ and $Y$ are independent, with

$$X \sim N(\mu, 1) \quad \text{and} \quad Y \sim N(\mu, 2).$$ \hfill (14.10)

Then two estimators are

$$\delta(x, y) = \frac{x + y}{2} \quad \text{and} \quad \delta^*(x, y) = \frac{2x + y}{3}.$$ \hfill (14.11)

Both are unbiased, and hence their risks (using squared-error loss) are their variances:

$$R(\mu; \delta) = \frac{1 + 2}{4} = \frac{3}{4} \quad \text{and} \quad R(\mu; \delta^*) = \frac{4 + 2}{9} = \frac{2}{3}.$$ \hfill (14.12)
Thus $\delta^*$ dominates $\delta$. (The second inequality in (14.9) holds for all $\mu$, although it does not have to for domination.) The concept of admissibility is based on lack of domination.

**Definition 24.** Let $\mathcal{D}$ be a set of decision procedures. A $\delta \in \mathcal{D}$ is **inadmissible** among procedures in $\mathcal{D}$ if there is another $\delta^* \in \mathcal{D}$ that dominates $\delta$. If there is no such $\delta^*$, then $\delta$ is **admissible** among procedures in $\mathcal{D}$.

If $\mathcal{D}$ is the set of unbiased estimators, then the UMVUE is admissible in $\mathcal{D}$, and any other unbiased estimator is inadmissible. Similarly, if $\mathcal{D}$ is the set of shift-equivariant procedures (assuming that restriction makes sense for the model), then the best shift-invariant estimator is the only admissible estimator in $\mathcal{D}$. In the Binomial example above, if $\mathcal{D}$ consists of the four given estimators, then all four are admissible in $\mathcal{D}$.

The presumption is that one would not want to use an inadmissible procedure, at least if risk is the only consideration. Other considerations, such as intuitivity or computational ease, may lead one to use an inadmissible procedure, provided it cannot be dominated by much. Conversely, any admissible procedure is presumed to be at least plausible, although there are some strange ones.
Chapter 14. The Decision-theoretic Approach

14.3.1 Stein’s surprising result

Admissibility is generally considered a fairly weak criterion. An admissible procedure does not have to be very good everywhere, but just have something going for it. Thus the statistical community was rocked when in 1956, Charles Stein showed that in the multivariate normal case, the usual estimator of the mean could be inadmissible\(^1\).

The model has random vector
\[ \mathbf{X} \sim N_p(\mu, I_p), \] (14.13)
with \( \mu \in \mathbb{R}^p \). The objective is to estimate \( \mu \) with squared-error loss, which in this case is multivariate:
\[ L(\mathbf{a}, \mu) = \| \mathbf{a} - \mu \|_2^2 = \sum_{i=1}^{p} (a_i - \mu_i)^2. \] (14.14)

The obvious estimator, which is the UMVUE, MLE, best shift-equivariant estimator, etc., is
\[ \delta(\mathbf{x}) = \mathbf{x}. \] (14.15)

Its risk is
\[ R(\mu; \delta) = E_\mu[\sum_{i=1}^{p} (X_i - \mu_i)^2] = p, \] (14.16)
because the \( X_i \)’s all have variance 1. When \( p = 1 \) or 2, then \( \delta \) is admissible. The surprise is that when \( p \geq 3 \), \( \delta \) is inadmissible. The most famous estimator that dominates it is the 1961 James-Stein estimator\(^2\),
\[ \delta_{JS}(\mathbf{x}) = \left( 1 - \frac{p - 2}{\| \mathbf{x} \|_2^2} \right) \mathbf{x}. \] (14.17)

It is a shrinkage estimator, because it takes the usual estimator, and shrinks it (towards 0 in this case), at least when \( (p - 2)/\| \mathbf{x} \|_2^2 < 1 \). Throughout the sixties and seventies, there was a frenzy of work on various shrinkage estimators. They are still quite popular. The domination result is not restricted to normality. It is quite broad. The general notion of shrinkage is very important in machine learning, where better predictions are found by restraining estimators from becoming too large (also called “regularization”).

To find the risk function for the James-Stein estimator when \( p \geq 3 \), start by writing
\[
R(\mu; \delta_{JS}) = E_\mu[\| \delta_{JS}(\mathbf{X}) - \mu \|^2] \\
= E_\mu[\| \mathbf{X} - \mu - \frac{p - 2}{\| \mathbf{X} \|_2^2} \mathbf{X} \|^2] \\
= E_\mu[\| \mathbf{X} - \mu \|^2] + E_\mu[\frac{(p - 2)^2}{\| \mathbf{X} \|_2^2}] - 2(p - 2)E_\mu[\frac{\mathbf{X}'(\mathbf{X} - \mu)}{\| \mathbf{X} \|_2^2}]. (14.18)
\]
The first term we recognize as from (14.16) as \( p \). Consider the third term, where
\[
E_\mu[\frac{\mathbf{X}'(\mathbf{X} - \mu)}{\| \mathbf{X} \|_2^2}] = \sum_{i=1}^{p} E_\mu[\frac{X_i(X_i - \mu_i)}{\| \mathbf{X} \|_2^2}]. (14.19)
\]


We take each term in the summation separately. The first one can be written

\[
E_\mu \left[ \frac{X_i(X_i - \mu_1)}{\|X\|^2} \right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{x_1(x_1 - \mu_1)}{\|x\|^2} \phi_{\mu_1}(x_1)dx_1\phi_{\mu_2}(x_2)dx_2 \cdots \phi_{\mu_p}(x_p)dx_p,
\]

where \( \phi \) is the \( N(0, 1) \) pdf, so that

\[
\phi_{\mu_1}(x_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_1-\mu_1)^2}.
\]

Looking at the innermost integral, note that by (14.21),

\[
(x_1 - \mu_1)\phi_{\mu_1}(x_1) = -\frac{\partial}{\partial x_1}\phi_{\mu_1}(x_1),
\]

so that

\[
\int_{-\infty}^{\infty} \frac{x_1(x_1 - \mu_1)}{\|x\|^2} \phi_{\mu_1}(x_1)dx_1 = -\int_{-\infty}^{\infty} \frac{x_1}{\|x\|^2} \left[ \frac{\partial}{\partial x_1}\phi_{\mu_1}(x_1) \right]dx_1.
\]

Now, with fixed \( x_2, \ldots, x_p \), use integration by parts:

\[
u = \frac{x_1}{\|x\|^2} \quad ; \quad u = \phi_{\mu_1}(x_1)
\]

\[
\frac{du}{dx_1} = \frac{1}{\|x\|^2} \frac{2x^2_1}{\|x\|^4} \quad ; \quad \frac{dv}{dx_1} = \frac{\partial}{\partial x_1}\phi_{\mu_1}(x_1)
\]

It is not hard to show that

\[
|u\nu|_{-\infty}^{\infty} = \frac{x_1}{\|x\|^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_1-\mu_1)^2} |_{-\infty}^{\infty} = 0.
\]

Then,

\[
\int_{-\infty}^{\infty} \frac{x_1(x_1 - \mu_1)}{\|x\|^2} \phi_{\mu_1}(x_1)dx_1 = -(u\nu)|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} vdu = \int_{-\infty}^{\infty} \left( \frac{1}{\|x\|^2} - \frac{2x^2_1}{\|x\|^4} \right) \phi_{\mu_1}(x_1)dx_1.
\]

Replacing the innermost integral in (14.20) with (14.26) yields

\[
E_\mu \left[ \frac{X_i(X_i - \mu_1)}{\|X\|^2} \right] = E_\mu \left[ \frac{1}{\|X\|^2} - \frac{2X^2_1}{\|X\|^4} \right].
\]

The same calculation works for \( i = 2, \ldots, p \), so that from (14.19),

\[
E_\mu \left[ \frac{X'(X - \mu)}{\|X\|^2} \right] = \sum_{i=1}^{p} E_\mu \left[ \frac{1}{\|X\|^2} - \frac{2X^2_i}{\|X\|^4} \right] = pE_\mu \left[ \frac{1}{\|X\|^2} - 2E_\mu \left[ \frac{\Sigma X^2}{\|X\|^2} \right] \right] = E_\mu \left[ p - \frac{2}{\|X\|^2} \right].
\]
Chapter 14. The Decision-theoretic Approach

Going back to (14.18),

\[ R(\mu; \delta_{JS}) = E_\mu[||X - \mu||^2] + E_\mu[\frac{(p-2)^2}{||X||^2}] - 2(p-2)E_\mu[\frac{X' (X - \mu)}{||X||^2}] \]

\[ = p + E_\mu[\frac{(p-2)^2}{||X||^2}] - 2(p-2)E_\mu[\frac{p-2}{||X||^2}] \]

\[ = p - E_\mu[\frac{(p-2)^2}{||X||^2}]. \] (14.29)

That’s it! The expected value at the end is positive, so that the risk is less than \( p \). That is

\[ R(\mu; \delta_{JS}) = p - E_\mu[\frac{(p-2)^2}{||X||^2}] < p = R(\mu; \delta) \] for all \( \mu \in \mathbb{R}^p \), (14.30)

meaning \( \delta(\bar{x}) = \bar{x} \) is inadmissible.

How much does the James-Stein estimator dominate \( \delta \)? It shrinks towards zero, so if the true mean is zero, one would expect the James-Stein estimator to be quite good. We can find the risk exactly, because when \( \mu = 0 \), \( ||X||^2 \sim \chi^2_p \), and it is not hard to show (using gamma’s) that \( E[1/\chi^2_p] = 1/(p-2) \), hence

\[ R(0; \delta_{JS}) = p - \frac{(p-2)^2}{p-2} = 2. \] (14.31)

Especially when \( p \) is large, this risk is much less than that of \( \delta \), which is always \( p \). Even for \( p = 3 \), the James-Stein risk is \( 2/3 \) of \( \delta \)’s. The farther from \( 0 \) the \( \mu \) is, the less advantage the James-Stein estimator has. As \( ||\mu|| \to \infty \), with \( ||X|| \sim \chi^2_p(||\mu||^2) \), the \( E_\mu[1/||X||^2] \to 0 \), so

\[ \lim_{||\mu|| \to \infty} R(\mu; \delta_{JS}) \to p = R(\mu; \delta). \] (14.32)

If rather than having good risk at zero, one has a “prior” idea that the mean is near some fixed \( \mu_0 \), one can instead shrink towards that vector:

\[ \delta^*_{JS}(\bar{x}) = (1 - \frac{p-2}{||\bar{x} - \mu_0||^2})(\bar{x} - \mu_0) + \mu_0. \] (14.33)

This estimator has the same risk as the regular James-Stein estimator, but with shifted parameter:

\[ R(\mu; \delta^*_{JS}) = p - E_\mu[\frac{(p-2)^2}{||\bar{x} - \mu_0||^2}] = p - E_\mu[\frac{(p-2)^2}{||\bar{x}||^2}], \] (14.34)

and has risk of 2 when \( \mu = \mu_0 \).

The James-Stein estimator itself is not admissible. There are many other similar estimators in the literature, some that dominate \( \delta_{JS} \) but are not admissible (such as the “positive part” estimator that does not allow the shrinking factor to be negative), and many admissible estimators that dominate \( \delta \).
14.4 Bayes procedures

One method for selecting among various $\delta$'s is to find one that minimizes the average of the risk, where the average is taken over $\Theta$. This averaging needs a probability measure on $\Theta$. From a Bayesian perspective, this distribution is the prior. From a frequentist perspective, it may or may not reflect prior belief, but it should be “reasonable.” The estimator that minimizes this average is the Bayes estimator corresponding to the distribution.

**Definition 25.** For given risk function $R(\theta;\delta)$ and distribution $\pi$ on $\Theta$, a Bayes procedure corresponding to $\pi$ is a procedure $\delta_\pi$ that minimizes the Bayes risk $E_\pi[R(\theta;\delta)]$ over $\delta$, i.e.,

$$E_\pi[R(\theta;\delta_\pi)] \leq E_\pi[R(\theta;\delta)] \text{ for any } \delta. \quad (14.35)$$

It might look daunting to minimize over an entire function, but we can reduce the problem to minimizing over a single value by using an iterative expected value. With both $X$ and $\theta$ random, the Bayes risk is the expected value of the loss over the joint distribution of $(X,\theta)$, hence can be written as the expected value of the conditional expected value of $L$ given $X$:

$$E_\pi[R(\theta;\delta)] = E[L(\delta(X),\theta)] = E[e_L(X)], \quad (14.36)$$

where

$$e_L(x) = E[L(\delta(x),\theta) | X = x]. \quad (14.37)$$

In that final expectation, $\theta$ is random, having the posterior distribution given $X = x$. In (14.37), because $x$ is fixed, $\delta(x)$ is just a constant, hence it may not be too difficult to minimize $e_L(x)$ over $\delta(x)$. Note that if we find the $\delta(x)$ to minimize $e_L(x)$ for each $x$, then we have also minimized the overall Bayes risk (14.36). Thus a Bayes procedure is $\delta_\pi$ such that

$$\delta_\pi(x) \text{ minimizes } E[L(\delta(x),\theta) | X = x] \text{ over } \delta(x) \text{ for each } x \in X. \quad (14.38)$$

In estimation with squared-error loss, (14.37) becomes

$$e_L(x) = E[(\delta(x) - g(\theta))^2 | X = x]. \quad (14.39)$$

That expression is minimized with $\delta(x)$ being the mean, in this case the conditional (posterior) mean of $g$:

$$\delta_\pi(x) = E[g(\theta) | X = x]. \quad (14.40)$$

To summarize:

**Lemma 19.** In estimating $g(\theta)$ with squared error loss, the Bayes procedure corresponding to the prior distribution $\pi$ is the posterior mean of $g(\theta)$.

A Bayesian does not care about $x$’s not observed, hence would immediately go to the conditional equation (14.38), and use the resulting $\delta_\pi(x)$. It is interesting that the decision-theoretic approach appears to bring Bayesians and frequentists together. They do end up with the same procedure, but from different perspectives. The Bayesian is trying to limit expected losses given the data, while the frequentist is trying to limit average expected losses, taking expected values as the experiment is repeated.
14.4.1 Bayes procedures and admissibility

A Bayes procedure with respect to a prior $\pi$ has good behavior averaging over the $\theta$. Any procedure that dominated that procedure would also have to have at least as good Bayes risk, hence would also be Bayes. In there is only one Bayes procedure for a $\pi$, then nothing could dominate it. Hence the following.

**Lemma 20.** Suppose $\delta_\pi$ is the unique Bayes procedure relative to the prior $\pi$. Then it is admissible.

**Proof.** Suppose $\delta^*$ dominates $\delta_\pi$. Then

$$R(\theta; \delta^*) \leq R(\theta; \delta_\pi) \quad \text{for all } \theta \in \Theta,$$

(14.41)

which implies that

$$\mathbb{E}_\pi[R(\theta; \delta^*)] \leq \mathbb{E}_\pi[R(\theta; \delta_\pi)].$$

(14.42)

But by definition, from (14.35),

$$\mathbb{E}_\pi[R(\theta; \delta_\pi)] \leq \mathbb{E}_\pi[R(\theta; \delta^*)].$$

(14.43)

Thus $\delta$ and $\delta_\pi$ are Bayes relative to $\pi$. But because the Bayes procedure is unique, $\delta^*$ and $\delta_\pi$ must be the same. Thus $\delta^*$ cannot dominate $\delta_\pi$, and $\delta_\pi$ is admissible. □

Not all admissible procedures are Bayes, but they are at least limits of Bayes procedures in some sense. Exactly which limits are admissible is a bit delicate, though. In any case, at least approximately, one can think of a procedure being admissible if there is some Bayesian somewhere, or a sequence of Bayesians, who would would use it. The book *Mathematical Statistics - A Decision Theoretic Approach* by Thomas Ferguson is an excellent introduction to these concepts.

14.5 Minimax procedures

Using a Bayes procedure still involves choosing a prior $\pi$, even if you are a frequentist and the prior does not really represent your beliefs. One attempt to objectifying the choice of a procedure is for each procedure, see what its worst risk is. Then you choose the procedure that has the best worst. For example, in Figure 14.1, the maximum risks are

$$
\begin{align*}
\text{MLE} & : 0.050 \\
 a = 1, b = 1 & : 0.026 \\
 a = 10, b = 10 & : 0.160 \\
 a = 10, b = 3 & : 0.346
\end{align*}
$$

(14.44)

Of these, the posterior mean for prior Beta$(1, 1)$ has the lowest maximum. It is called the minimax procedure.

**Definition 26.** Let $\mathcal{D}$ be a set of decision procedures. A $\delta \in \mathcal{D}$ is minimax among procedures in $\mathcal{D}$ if for any other $\delta^* \in \mathcal{D}$,

$$
\sup_{\theta \in \Theta} R(\theta; \delta) \leq \sup_{\theta \in \Theta} R(\theta; \delta^*).$$

(14.45)
For estimating the normal mean as in Section 14.3.1, the usual estimator \( \delta(x) = x \) is minimax, as is any estimator (such as the James-Stein estimator) that dominates it. In particular, we see that minimax procedures do not have to be unique.

Again looking at Figure 14.1, note that the minimax procedure is the flattest. In fact, it is also “maximin” in that it has the worst best risk. It looks as if when trying to limit bad risk everywhere, you give up very good risk somewhere. This idea leads to one method for finding a minimax procedure: A Bayes procedure with flat risk is minimax.

**Lemma 21.** Suppose \( \delta_\pi \) is Bayes with respect to the prior \( \pi \), and the risk is constant,

\[
R(\theta; \delta_\pi) = c \text{ for all } \theta \in \Theta.
\]  

(14.46)

Then \( \delta_\pi \) is minimax.

*Proof.* Suppose \( \delta_\pi \) is not minimax, so that there is a \( \delta^* \) such that

\[
\sup_{\theta \in \Theta} R(\theta; \delta^*) < \sup_{\theta \in \Theta} R(\theta; \delta_\pi) = c.
\]  

(14.47)

But then

\[
E_\pi[R(\theta; \delta^*)] < c = E_\pi[R(\theta; \delta_\pi)],
\]  

(14.48)

meaning that \( \delta_\pi \) is not Bayes. Hence we have a contradiction, so that \( \delta_\pi \) is minimax. \( \Box \)

**14.5.1 Example: Binomial**

Suppose \( X \sim Binomial(n, \theta) \), \( \theta \in (0, 1) \), then from (10.45), the MSE for the Bayes estimator using the Beta(\( \alpha \), \( \beta \)) prior can be calculated as

\[
R(\theta; \delta_{\alpha, \beta}) = \frac{n \theta (1 - \theta)}{(n + \alpha + \beta)^2} + \left( \frac{\theta - \frac{\alpha}{n + \alpha + \beta}}{n + \alpha + \beta} \right)^2.
\]  

(14.49)

If we can find an \((\alpha, \beta)\) so that this risk is constant, then the corresponding estimator is minimax.
Chapter 15

Asymptotics: Convergence in probability

Often, inference about a single function of the parameter $g(\theta)$ consists of an estimate $\hat{g}(\theta)$ plus an assessment of its variability, which may be in the form of a standard error of the estimate, a confidence or probability interval for $g(\theta)$, or a significance level or probability of a particular hypothesized value of $g(\theta)$. In any case, we need more than just the estimate.

The standard error of an estimate $\hat{g}(\theta)$ is its standard deviation or, in a slight ambiguity of notation, an estimate of such:

$$SE[\hat{g}(\theta)] = \sqrt{Var[\hat{g}(\theta)]} \quad \text{or} \quad \hat{\sqrt{Var[\hat{g}(\theta)]}}.$$ (15.1)

Then an approximate 95% confidence interval for $g(\theta)$ would be

$$\hat{g}(\theta) \pm 2 \cdot SE[\hat{g}(\theta)].$$ (15.2)

If the estimator is Normal with mean $g(\theta)$ and variance $SE^2$, then this interval is exact if “2” is changed to “1.96.” For example, when estimating the mean $\mu$ of a normal when $\sigma^2$ is known, $X \pm 1.96\sigma/\sqrt{n}$ is an exact 95% confidence interval. In other situations, (15.2) may or not be a good approximation. Unfortunately, once one moves away from linear estimators from Normal data, it is typically very difficult to figure out exactly the distribution of an estimator, or even what its mean and variance are. We have been in this situation many times, with MLE’s, Pitman Estimators, even the median.

One way to address such questions is to look at what happens when $n$ is large, or actually, as $n \to \infty$. In many case, nice asymptotic results are available, and they give surprisingly good approximations even when $n$ is nowhere near $\infty$.

15.1 The set-up

We assume that we have a sequence of random variables, or random vectors. That is, for each $n$, we have a random $p \times 1$ vector $W_n$ with space $W_n(\subset \mathbb{R}^p)$ and probability distribution $P_n$. There need not be any particular relationship between the $W_n$'s for different $n$'s, but in the most common situation we will deal with, $W_n$ is some function of iid $X_1, \ldots, X_n$, so as $n \to \infty$, the function is based on more observations.
For example, \( W_n \) may be the sample mean of \( X_1, \ldots, X_n \), iid \( N(\mu, \sigma^2) \)'s, so that
\[
W_n = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n}).
\] (15.3)

Or \( W_n \) could be a vector,
\[
W_n = \left( \frac{\bar{X}_n}{S_n^2} \right),
\] (15.4)
whose joint distribution we know. In other situations, the parameters may depend on \( n \), e.g.,
\[
W_n \sim \text{Binomial}(n, \frac{\lambda}{n})
\] (15.5)
for some \( \lambda > 0 \). (Technically, we would have to set the \( p \) to be \( \min\{1, \lambda/n\} \), but as \( n \to \infty \), eventually that would be \( \lambda/n \).)

The two types of convergence we will consider are convergence in probability to a constant and convergence in distribution to a random vector, which will be defined in the next two sections.

### 15.2 Convergence in probability to a constant

A sequence of constants \( a_n \) approaching the constant \( c \) means that as \( n \to \infty \), \( a_n \) gets arbitrarily close to \( c \); technically, for any \( \epsilon > 0 \), eventually \( |a_n - c| < \epsilon \). That definition does not immediately transfer to random variables. For example, suppose \( \bar{X}_n \) is the mean of \( n \) iid \( N(\mu, \sigma^2) \)'s as in (15.3). The law of large numbers says that as \( n \to \infty \), \( \bar{X} \to \mu \). But that cannot be always true, since no matter how large \( n \) is, the space of \( \bar{X}_n \) is \( \mathbb{R} \). On the other hand, the probability is high that \( \bar{X}_n \) is close to \( \mu \). That is, for any \( \epsilon > 0 \),
\[
P_n[|\bar{X}_n - \mu| < \epsilon] = P[|N(0,1)| < \sqrt{n}\epsilon/\sigma] = \Phi(\sqrt{n}\epsilon/\sigma) - \Phi(-\sqrt{n}\epsilon/\sigma),
\] (15.6)
where \( \Phi \) is the distribution function of \( Z \sim N(0,1) \). Now let \( n \to \infty \). Because \( \Phi \) is a distribution function, the first \( \Phi \) on the right in (15.6) goes to 1, and the second goes to 0, so that
\[
P_n[|\bar{X}_n - \mu| < \epsilon] \to 1.
\] (15.7)
Thus \( \bar{X}_n \) isn’t for sure close to \( \mu \), but is with probability 0.9999999999 (assuming \( n \) is large enough). Now for the definition.

**Definition 27.** The sequence of random variables \( W_n \) **converges in probability** to the constant \( c \), written
\[
W_n \to^P c,
\] (15.8)
if for every \( \epsilon > 0 \),
\[
P_n[|W_n - c| < \epsilon] \to 1.
\] (15.9)

If \( W_n \) is a sequence of random \( p \times 1 \) vectors, and \( \zeta \) is a \( p \times 1 \) constant vector, then \( W_n \to^P \zeta \) if for every \( \epsilon > 0 \),
\[
P_n[||W_n - \zeta|| < \epsilon] \to 1.
\] (15.10)
It turns out that $W_n \to^P c$ if and only if each component $W_{ni} \to^P c_i$, where $W_n = (W_{n1}, \ldots, W_{np})$.

As an example, suppose $X_1, \ldots, X_n$ are iid Uniform(0, 1), and let $W_n = \min\{X_1, \ldots, X_n\}$. You would expect that as the number of observations between 0 and 1 increase, the minimum would get pushed down to 0. So the question is whether $W_n \to^P 0$. To prove it, take any $1 > \epsilon > 0$, and look at

$$P_n[|W_n - 0| < \epsilon] = P_n[|W_n| < \epsilon] = P_n[W_n < \epsilon],$$

(15.11)

because $W_n$ is positive. Then, recalling the pdf of the minimum of uniforms in (6.96, with $k = 1$),

$$P_n[W_n < \epsilon] = \int_0^\epsilon 1 - (1 - z)^n dz = - (1 - z)^n \bigg|_0^\epsilon = 1 - (1 - \epsilon)^n \to 1 \text{ as } n \to \infty.$$ (15.12)

Thus

$$\min\{X_1, \ldots, X_n\} \to^P 0.$$ (15.13)

### 15.2.1 Chebychev’s Inequality

The examples in (15.6) and (15.12) are unusual in that we can calculate the probabilities exactly. It is more common that some inequalities are used. Chebychev’s, for instance, which can be used to show that $W_n \to^P 0$ if $E[(W_n - c)^2] \to 0$, comes next.

**Lemma 22.** **Chebychev’s Inequality.** For random variable $W$ and $\epsilon > 0$,

$$P[|W| \geq \epsilon] \leq \frac{E[W^2]}{\epsilon^2}.$$ (15.14)

**Proof.**

$$E[W^2] = E[W^2I_{|W|<\epsilon}] + E[W^2I_{|W|\geq\epsilon}] \geq E[W^2I_{|W|\geq\epsilon}] \geq \epsilon^2 E[I_{|W|\geq\epsilon}] = \epsilon^2 P[|W| \geq \epsilon].$$ (15.15)

Then (15.14) follows. QED

**Aside 1.** A similar proof can be applied to any nondecreasing function $\phi(w) : [0, \infty) \to \mathbb{R}$ to show that

$$P[|W| \geq \epsilon] \leq \frac{E[\phi(|W|)]}{\phi(\epsilon)}.$$ (15.16)

Chebychev’s Inequality uses $\phi(w) = w^2$. The general form is called **Markov’s Inequality**.

\footnote{Why is it ok for us to ignore $\epsilon \geq 1$?}
Chapter 15. Asymptotics: Convergence in probability

Now apply Chebychev to the sequence $W_n$, substituting $W = W_n - c$, so that for any $\epsilon > 0$, 
\[ P_n[|W_n - c| \geq \epsilon] \leq \frac{E[(W_n - c)^2]}{\epsilon^2}. \] (15.17)

Then 
\[ E[(W_n - c)^2] \to 0 \implies P_n[|W_n - c| < \epsilon] > 1 - \frac{E[(W_n - c)^2]}{\epsilon^2} \to 1. \] (15.18)

A couple of useful consequences:

1. If $\delta_n(x_1, \ldots, x_n)$ is an estimate of $g(\theta)$ for every $n$, and if the mean square error goes to 0 as $n \to \infty$, then 
\[ \delta_n(X_1, \ldots, X_n) \to^P g(\theta). \] (15.19)

(That is, $W_n = \delta_n(X_1, \ldots, X_n)$ and $c = g(\theta)$ in (15.18).) A sequence of estimators for which (15.19) holds is said to be consistent.

**Definition 28.** Suppose $\delta_1, \delta_2, \ldots$ is a sequence of estimators of $g(\theta)$. (That is, for each $n$, there is a set of probability measures $\{P_{n\theta} \mid \theta \in \Theta\}$.) Then if 
\[ \delta_n \to^P g(\theta) \] (15.20)

for each $\theta \in \Theta$, the sequence of estimators is consistent.

Notice that this definition assumes the same parameter space for each $n$. One often leaves out the words “sequence of”, and says that “$\delta_n$ is a consistent estimator of $g(\theta)$.”

2. If $X_1, \ldots, X_n$ are iid with mean $\mu$ and variance $\sigma^2 < \infty$, then 
\[ \overline{X}_n \to^P \mu. \] (15.21)

This fact follows from part 1, since the mean square error of $\overline{X}_n$ as an estimator of $\mu$ is $Var[\overline{X}_n] = \sigma^2/n \to 0$. Equation (15.21) is called the weak law of large numbers.

Revisiting the uniform example, suppose that $X_1, \ldots, X_n$ are iid $Uniform(\mu, \mu + 1)$, and consider $\delta_n(x_1, \ldots, x_n) = \min\{x_1, \ldots, x_n\}$ to be an estimator of $\mu$. Because $\delta_n$ is shift-equivariant, we know that the mean square error is the same for all $\mu$. Thus by (6.96), $\delta_n \sim Beta(1, n)$ if $\mu = 0$, and 
\[ MSE(\delta_n) = Var_0[\delta_n] + E_0[\delta_n]^2 = \frac{n}{(n+1)^2(n+2)} + \frac{1}{(n+1)^2} \to 0. \] (15.22)

Hence $\min\{X_1, \ldots, X_n\} \to^P \mu$, i.e., the minimum is a consistent estimator of $\mu$.

The next section contains some useful techniques for showing convergence in probability.
15.3 The law of large numbers; Mapping

Showing an estimator is consistent usually begins with the law of large numbers. Here is the formal statement.

**Lemma 23. Weak law of large numbers (WLLN).** If \( X_1, \ldots, X_n \) are iid with (finite) mean \( \mu \), then

\[
\overline{X}_n \longrightarrow^P \mu. \tag{15.23}
\]

The difference between this lemma and what we stated in (15.21) is that before, we assumed that the variance was finite. We had to do that to use Chebychev’s inequality. It turns out that the result still holds as long as the mean is finite. We will use that result, but not prove it. It can be applied to means of functions of the \( X_i \)’s. For example, if \( E[X_i^2] < \infty \), then

\[
\frac{1}{n} \sum_{i=1}^{n} X_i^2 \longrightarrow^P E[X_i]^2 = \mu^2 + \sigma^2, \tag{15.24}
\]

because the \( X_1^2, \ldots, X_n^2 \) are iid with mean \( \mu^2 + \sigma^2 \).

Not only means of functions, but functions of the mean are of interest. For example, if the \( X_i \)’s are iid Exponential(\( \lambda \)), then the mean is \( 1/\lambda \), so that

\[
\frac{1}{\overline{X}_n} \longrightarrow^P \frac{1}{\lambda}. \tag{15.25}
\]

But we really want to estimate \( \lambda \). The method of moments and maximum likelihood estimators are both \( 1/\overline{X}_n \), though not the UMVUE. Does

\[
\frac{1}{\overline{X}_n} \longrightarrow^P \lambda? \tag{15.26}
\]

We could find the mean and variance of \( 1/\overline{X}_n \), but more simply we note that if \( \overline{X}_n \) is close to \( 1/\lambda \), \( 1/\overline{X}_n \) must be close to \( \lambda \), because the function \( 1/w \) is continuous. Formally, we have the following mapping result.

**Lemma 24.** If \( W_n \longrightarrow^P c \), and \( g(w) \) is a function continuous at \( w = c \), then

\[
g(W_n) \longrightarrow^P g(c). \tag{15.27}
\]

**Proof:** By definition of continuity, for every \( \epsilon > 0 \), there exists a \( \delta > 0 \) such that

\[
|w - c| < \delta \implies |g(w) - g(c)| < \epsilon. \tag{15.28}
\]

Thus the event on the right happens at least as often as that on the left, i.e.,

\[
P_n[|W_n - c| < \delta] \leq P_n[|g(W_n) - g(c)| < \epsilon]. \tag{15.29}
\]

The definition of \( \longrightarrow^P \) means that \( P_n[|W_n - c| < \delta] \rightarrow 1 \) for any \( \delta > 0 \), but \( P_n[|g(W_n) - g(c)| < \epsilon] \) is larger, hence

\[
P_n[|g(W_n) - g(c)| < \epsilon] \longrightarrow 1, \tag{15.30}
\]

proving (15.27).

Thus the answer to (15.26) is “Yes.” This lemma also works for vector \( W_n \), that is, if \( g(w) \) is continuous at \( c \), then

\[
W_n \longrightarrow^P c \implies g(W_n) \longrightarrow^P g(c). \tag{15.31}
\]
Chapter 15. Asymptotics: Convergence in probability

Examples.

1. Suppose $X_1, \ldots, X_n$ are iid with exponential family density, where $\theta$ is the natural parameter and $x$ is the natural statistic:

$$f(x; \theta) = a(x)e^{\theta x - \psi(\theta)}.$$  \hspace{1cm} (15.32)

Then the WLLN shows that

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to^P E[X_i] = \mu(\theta) = \psi'(\theta).$$  \hspace{1cm} (15.33)

The MLE is the solution to the equation $x_n = \mu(\theta)$ (why?), hence is

$$\hat{\theta}_n = \mu^{-1}(x_n).$$ \hspace{1cm} (15.34)

The function $\mu^{-1}(w)$ is continuous, hence by Lemma 24,

$$\hat{\theta}_n = \mu^{-1}(x_n) \to^P \mu^{-1}(\mu(\theta)) = \theta.$$ \hspace{1cm} (15.35)

Thus the MLE is a consistent estimator of $\theta$.

2. Consider regression through the origin, that is, $(X_1, Y_2), \ldots, (X_n, Y_n)$ are iid,

$$E[Y_i | X_i = x_i] = \beta x_i, \quad Var(Y_i | X_i = x_i) = \sigma^2, \quad E[X_i] = \mu_X, \quad Var(X_i) = \sigma^2_X > 0.$$  \hspace{1cm} (15.36)

The least squares estimate of $\beta$ is

$$\hat{\beta}_n = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$ \hspace{1cm} (15.37)

Is this a consistent estimator? We know from the WLLN by (15.24) that

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 \to \mu^2_X + \sigma^2_X.$$ \hspace{1cm} (15.38)

Also, $X_1 Y_1, \ldots, X_n Y_n$ are iid, and

$$E[X_i Y_i | X_i = x_i] = x_i E[Y_i | X_i = x_i] = \beta x_i^2,$$ \hspace{1cm} (15.39)

hence

$$E[X_i Y_i] = E[\beta X_i^2] = \beta(\mu^2_X + \sigma^2_X).$$ \hspace{1cm} (15.40)

Thus the WLLN shows that

$$\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \to^P \beta(\mu^2_X + \sigma^2_X).$$ \hspace{1cm} (15.41)

Now consider

$$W_n = \left( \frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right) \quad \text{and} \quad \xi = \left( \begin{array}{c} \beta(\mu^2_X + \sigma^2_X) \\ \mu^2_X + \sigma^2_X \end{array} \right),$$ \hspace{1cm} (15.42)
so that $W_n \to^P c$. The function $g(w_1, w_2) = w_1 / w_2$ is continuous at $w = c$, hence (15.31) shows that
\[
g(W_n) = \frac{1}{n} \sum_{i=1}^{n} X_i Y_i \to^P g(c) = \frac{\beta (\mu_X + \sigma_X^2)}{\mu_X^2 + \sigma_X^2},
\]
that is,
\[
\hat{\beta}_n = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \to^P \beta.
\]
So, yes, the least squares estimator is consistent.
Convergence to a constant is helpful, but generally more information is needed, as for confidence intervals. E.g., is we can say that
\[
\frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \approx N(0, 1),
\]
then an approximate 95% confidence interval for \(\theta\) would be
\[
\hat{\theta} \pm 2 \times SE(\hat{\theta}),
\]
where the “2” is approximately 1.96. Thus we need to find the approximate distribution of a random variable. In the asymptotic setup, we need the notion of \(W_n\) converging to a random variable. It is formalized by looking at the respective distribution function for each possible value, almost. Here is the definition.

**Definition 29.** Suppose \(W_n\) is a sequence of random variables, and \(W\) is a random variable. Let \(F_n\) be the distribution function of \(W_n\), and \(F\) be the distribution function of \(W\). Then \(W_n\) converges in distribution to \(W\) if
\[
F_n(w) \rightarrow F(w)
\]
for every \(w \in \mathbb{R}\) at which \(F\) is continuous. This convergence is written
\[
W_n \rightarrow^D W.
\]

For example, go back to \(X_1, \ldots, X_n\) iid \(Uniform(0, 1)\), but now let
\[
W_n = n \min\{X_1, \ldots, X_n\}.
\]
(The minimum itself goes to 0, but by multiplying by \(n\) it may not.) The distribution function of \(W_n\) is \(F_n(w) = 0\) if \(w \leq 0\), and if \(w > 0\) (using calculations as in equation 15.12),
\[
F_n(w) = P[W_n \leq w] = P[\min\{X_1, \ldots, X_n\} \leq w/n] = 1 - \left(1 - \frac{w}{n}\right)^n.
\]
Now let $n \to \infty$. Recalling that $(1 + z/n)^n \to e^z$,

$$F_n(w) \to \begin{cases} 1 - e^{-w} & \text{if } w > 0 \\ 0 & \text{if } w \leq 0. \end{cases}$$

(16.7)

Is the right hand side a distribution function of some random variable? Yes, indeed, it is the $F$ of $W \sim \text{Exponential}(1)$. (This $F$ is continuous, so one can switch the equality from the $\leq$ to the $>$.) That is,

$$n \min\{X_1, \ldots, X_n\} \to^D \text{Exponential}(1).$$

(16.8)

**Aside 2.** In the definition, the convergence (16.3) does not need to hold at $w$’s for which $F$ is not continuous. This relaxation exists because sometimes the limit of the $F_n$’s will have points at which the function is continuous from the left but not the right, whereas $F$’s need to be continuous from the right. For example, take $W_n = \text{Bernoulli}(1/2) + 1/n$. It seems reasonable that $W_n \to^D \text{Bernoulli}(1/2)$. Let $F_n$ be the distribution function for $W_n$:

$$F_n(w) = \begin{cases} 0 & \text{if } w < 1/n \\ 1/2 & \text{if } 1/n \leq w < 1 + 1/n \\ 1 & \text{if } 1 + 1/n \leq w. \end{cases}$$

(16.9)

Now let $n \to \infty$, so that

$$F_n(w) \to \begin{cases} 0 & \text{if } w \leq 0 \\ 1/2 & \text{if } 0 < w \leq 1 \\ 1 & \text{if } 1 < w. \end{cases}$$

(16.10)

That limit is **not** a distribution function, though it would be if the $\leq$’s and $<$’s were switched, in which case it would be the $F$ for a Bernoulli(1/2). Luckily, the definition allows the limit to be wrong at points of discontinuity, which are 0 and 1 in this example, so we **can** say that $W_n \to^D \text{Bernoulli}(1/2)$.

For another example, suppose $X_n \sim \text{Binomial}(n, \lambda/n)$ for some fixed $\lambda > 0$. The distribution function of $X_n$ is

$$F_n(x) = \begin{cases} 0 & \text{if } x < 0 \\ \sum_{i=0}^{\lfloor x \rfloor} f(i; n, \lambda/n) & \text{if } 0 \leq x \leq n \\ 1 & \text{if } n < x. \end{cases}$$

(16.11)

where $\lfloor x \rfloor$ is the largest integer less than or equal to $x$, and $f(i; n, \lambda/n)$ is the Binomial $(n, \lambda/n)$ pmf. Taking the limit as $n \to \infty$ of $F_n$ requires taking the limit of the $f$’s, so we will do that first. For a positive integer $i \leq n$,

$$f(i; n, \lambda/n) = \frac{n!}{i!(n-i)!} \left( \frac{\lambda}{n} \right)^i \left( 1 - \frac{\lambda}{n} \right)^{n-i} = \frac{\lambda^i}{i!} \frac{n(n-1) \cdots (n-i+1)}{n^i} \left( \frac{1 - \lambda}{n} \right)^n \left( 1 - \frac{\lambda}{n} \right)^{-i}. \quad (16.12)$$

Now $i$ is fixed and $n \to \infty$. Consider the various factors on the right. The first has no $n$’s. The second has $i$ terms on the top, and $n^i$ on the bottom, so can be written

$$\frac{n(n-1) \cdots (n-i+1)}{n^i} = \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-i+1}{n} = 1 \frac{1}{n} \cdots \frac{1-i}{n} \to 1. \quad (16.13)$$
The third term goes to $e^{-\lambda}$, and the fourth goes to 1. Thus for any positive integer $i$,

$$f(i; n, \frac{\lambda}{n}) \longrightarrow e^{-\lambda} \frac{\lambda^i}{i!}$$

(16.14)

which is the pmf of the Poisson($\lambda$). Going back to the $F_n$ in (16.11), note that no matter how large $x$ is, as $n \to \infty$, eventually $x < n$, so that the third line never comes into play. Thus

$$F_n(x) \longrightarrow \begin{cases} 
0 & \text{if } x < 0 \\
\sum_{i=0}^{|x|} e^{-\lambda} \frac{\lambda^i}{i!} & \text{if } 0 \leq x.
\end{cases}$$

(16.15)

But that is the distribution function of the Poisson($\lambda$), i.e.,

$$\text{Binomial}(n, \frac{\lambda}{n}) \longrightarrow^D \text{Poisson}(\lambda).$$

(16.16)

## 16.1 Converging to a constant random variable

It could be that $W_n \longrightarrow^D W$, where $P[W = c] = 1$, that is, $W$ is a constant random variable. (Sounds like an oxymoron.) But that looks like $W_n$ is converging to a constant. It is. In fact,

$$W_n \longrightarrow^D W \text{ where } P[W = c] \text{ if and only if } W_n \longrightarrow^P c.$$  

(16.17)

Let $F_n(w)$ be the distribution function of $W_n$, and $F$ the distribution function of $W$, so that

$$F(w) = \begin{cases} 
0 & \text{if } w < c \\
1 & \text{if } c \leq w.
\end{cases}$$

(16.18)

For any $\epsilon > 0$,

$$P[|W_n - c| \leq \epsilon] = P[W_n \leq c + \epsilon] - P[W_n < c - \epsilon].$$

(16.19)

Also,

$$P[W_n \leq c - 3\epsilon/2] \leq P[W_n < c - \epsilon] \leq P[W_n \leq c - \epsilon/2],$$

(16.20)

hence, because $F_n(w) = P[W_n \leq w],

$$F_n(c + \epsilon) - F_n(c - \epsilon/2) \leq P[|W_n - c| \leq \epsilon] \leq F_n(c + \epsilon) - F_n(c - 3\epsilon/2).$$

(16.21)

We use this equation to show (16.17).

1. First suppose that $W_n \longrightarrow^D W$, so that $F_n(w) \to F(w)$ if $w \neq c$. Then applying the convergence to $w = c + \epsilon$ and $c - \epsilon/2$,

$$F_n(c + \epsilon) \to F(c + \epsilon) = 1 \text{ and } F_n(c - \epsilon/2) \to F(c - \epsilon/2) = 0.$$  

(16.22)

Then (16.21) shows that $P[|W_n - c| \leq \epsilon] \to 1$, proving that $W_n \longrightarrow^P c$.

2. Next, suppose $W_n \longrightarrow^P c$. Then $P[|W_n - c| \leq \epsilon] \to 1$, hence from (16.21),

$$F_n(c + \epsilon) - F_n(c - 3\epsilon/2) \longrightarrow 1.$$  

(16.23)

The only way that can happen is for

$$F_n(c + \epsilon) \to 1 \text{ and } F_n(c - 3\epsilon/2) \to 0.$$  

(16.24)

But since that holds for any $\epsilon > 0$, for any $w < c$, $F_n(w) \to 0$, and for any $w > c$, $F_n(w) \to 1$. That is, $F_n(w) \to F(w)$ for $w \neq c$, proving that $W_n \longrightarrow^D W$. 

16.2 Moment generating functions

It may be difficult to find distribution functions and their limits, but often moment generating functions are easier to work with. Recall that if two random variables have the same moment generating function that is finite in a neighborhood of 0, then they have the same distribution. It is true also of limits, that is, if the mgf’s of a sequence of random variables converge to a mgf, then the random variable converge.

Lemma 25. Suppose $W_1, W_2, \ldots$ is a sequence of random variables, where $M_n(t)$ is the mgf of $W_n$, and suppose $W$ is a random variable with mgf $M(t)$. If for some $\varepsilon > 0, M_n(t) < \infty$ for all $n$ and $M(t) < \infty$ for all $|t| < \varepsilon$,

\[ W_n \xrightarrow{D} W \text{ if and only if } M_n(t) \xrightarrow{} M(t) \text{ for all } |t| < \varepsilon. \quad (16.25) \]

Looking again at the binomial example, if $X_n \sim \text{Binomial}(n, \lambda/n)$, then its mgf is

\[ M_n(t) = ((1 - \frac{\lambda}{n}) + \frac{\lambda}{n} e^t)^n = (1 + \frac{-\lambda + \lambda e^t}{n})^n. \quad (16.26) \]

Letting $n \to \infty$,

\[ M_n(t) \xrightarrow{} e^{-\lambda} + \lambda e^t, \quad (16.27) \]

which is the mgf of a Poisson($\lambda$), hence (16.16).

16.3 The Central Limit Theorem

We know that sample means tend to the population mean, if the latter exists. But we can obtain more information with a distributional limit. In the normal case, we know that the sample mean is normal, and with appropriate normalization, it is $\text{Normal}(0,1)$. A Central Limit Theorem is one that says a properly normalized sample mean approaches normality even if the original variables do not.

Start with $X_1, \ldots, X_n$ iid with mean 0 and variance 1, and mgf $M_X(t)$, which is finite for $|t| < \varepsilon$ for some $\varepsilon > 0$. The variance of $\overline{X}_n$ is $1/n$, so to normalize it we multiply by $\sqrt{n}$:

\[ W_n = \sqrt{n} \overline{X}_n. \quad (16.28) \]

To find the asymptotic distribution of $W_n$, we first find its mgf, $M_n(t)$:

\[
M_n(t) = E[e^{tW_n}] = E[e^{t\sqrt{n} \overline{X}_n}] = E[e^{(t/\sqrt{n}) \sum X_i}] = E[e^{(t/\sqrt{n}) X_i^n}] = M_X(t/\sqrt{n}). \quad (16.29)
\]

Now $M_n(t)$ is finite if $M_X(t/\sqrt{n})$ is, and $M_X(t/\sqrt{n}) < \infty$ if $|t/\sqrt{n}| < \varepsilon$, which is certainly true if $|t| < \varepsilon$. That is, $M_n(t) < \infty$ if $|t| < \varepsilon$.

To find the limit of $M_n$, first, take logs:

\[
\log(M_n(t)) = n \log(M_X(t/\sqrt{n})) = n R_X(t/\sqrt{n}), \quad (16.30)
\]

where $R_X(t) = \log(M_X(t))$. Expand $R_X$ in a Taylor Series about $t = 0$:

\[
R_X(t) = R_X(0) + t R_X'(0) + \frac{t^2}{2} R''(t^*), \quad t^* \text{ between 0 and t}. \quad (16.31)
\]
16.3. The Central Limit Theorem

But \( R_X(0) = 0 \), and \( R'_X(0) = E[X] = 0 \), by assumption. Thus substituting \( t/\sqrt{n} \) for \( t \) in (16.31) yields

\[
R_X(t/\sqrt{n}) = \frac{t^2}{2n} R''(t_n^*), \quad t_n^* \text{ between 0 and } t/\sqrt{n},
\]

(16.32)
hence by (16.30),

\[
\log(M_n(t)) = \frac{t^2}{2} R''(t_n^*).
\]

(16.33)
The mgf \( M_X(t) \) has all its derivatives as long as \( |t| < \epsilon \), which means so does \( R_X \). In particular, \( R''(t) \) is continuous at \( t = 0 \). As \( n \to \infty \), \( t_n^* \) gets squeezed between 0 and \( t/\sqrt{n} \), hence \( t_n^* \to 0 \), and

\[
\log(M_n(t)) = \frac{t^2}{2} R''(t_n^*) \to \frac{t^2}{2} R''(0) = \frac{t^2}{2} \text{Var}[X_i] = \frac{t^2}{2},
\]

(16.34)
because we have assumed that \( \text{Var}[X_i] = 1 \). Finally,

\[
M_n(t) \to e^{t^2/2},
\]

(16.35)
which is the mgf of a \( \text{N}(0,1) \), i.e.,

\[
\sqrt{n} \overline{X_n} \to^D N(0,1).
\]

(16.36)

There are many central limit theorems, depending on various assumptions, but the most basic is the following.

**Theorem 10. Central limit theorem.** Suppose \( X_1, X_2, \ldots \) are iid with mean 0 and variance 1. Then (16.36) holds.

What we proved using (16.35) required the mgf be finite in a neighborhood of 0. This theorem does not need mgf’s, only that the variance is finite. A slight generalization of the theorem has \( X_1, X_2, \ldots \) iid with mean \( \mu \) and variance \( \sigma^2 \), \( 0 < \sigma^2 < \infty \), and concludes that

\[
\sqrt{n} (\overline{X_n} - \mu) \to^D N(0,\sigma^2).
\]

(16.37)

### 16.3.1 Supersizing

Convergence in distribution immediately translates to multivariate random variables. That is, suppose \( W_n \) is a \( p \times 1 \) random vector with distribution function \( F_n \). Then

\[
W_n \to^D W
\]

(16.38)
for some \( p \times 1 \) random vector \( W \) with distribution function \( F \) if

\[
F_n(w) \to F(w)
\]

(16.39)
for all points \( w \in \mathbb{R} \) at which \( F(w) \) is continuous.

If \( M_n(t) \) is the mgf of \( W_n \), and \( M(t) \) is the mgf of \( W \), and these mgf’s are all finite for \( \|t\| < \epsilon \) for some \( \epsilon > 0 \), then

\[
W_n \to^D W \text{ iff } M_n(t) \to M(t) \text{ for all } \|t\| < \epsilon.
\]

(16.40)
Now for the central limit theorem. Suppose \( X_1, X_2, \ldots \) are iid random vectors with mean \( \mu \), covariance matrix \( \Sigma \), and mgf \( M_X(t) < \infty \) for \( \|t\| < \epsilon \). Let
\[
W_n = \sqrt{n}(\bar{X}_n - \mu).
\] (16.41)

Let \( \mathbf{a} \) be any \( p \times 1 \) vector, and write
\[
\mathbf{a}'W_n = \sqrt{n}(\mathbf{a}'X_n - \mathbf{a}'\mu) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbf{a}'X_i - \mathbf{a}'\mu \right).
\] (16.42)

Thus \( \mathbf{a}'W_n \) is the normalized sample mean of the \( \mathbf{a}'X_i \)'s, and the regular central limit theorem 10 (actually, equation 16.37) can be applied, where \( \sigma^2 = \text{Var}[\mathbf{a}'X_i] = \mathbf{a}'\Sigma \mathbf{a} \):
\[
\mathbf{a}'W_n \xrightarrow{D} N(0, \mathbf{a}'\Sigma \mathbf{a}).
\] (16.43)

But then that means the mgf’s converge in (16.43):
\[
E[e^{t(\mathbf{a}'W_n)}] \xrightarrow{\epsilon} e^{\frac{1}{2} t^2 \mathbf{a}'\Sigma \mathbf{a}},
\] (16.44)
for \( t\|\mathbf{a}\| < \epsilon \). Now switch notation so that \( \mathbf{a} = t \) and \( t = 1 \), and we have that
\[
M_n(t) = E[e^{t(\mathbf{a}'W_n)}] \xrightarrow{\epsilon} e^{\frac{1}{2} t^2 \mathbf{a}'\Sigma \mathbf{a}},
\] (16.45)

which holds for any \( \|t\| < \epsilon \). The right hand side is the mgf of a \( N(0, \Sigma) \), so
\[
W_n = \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \Sigma).
\] (16.46)

In fact, (16.46) holds whenever \( \Sigma \) is finite.

**Example.** Suppose \( X_1, X_2, \ldots \) are iid with mean \( \mu \) and variance \( \sigma^2 \). One might be interested in the joint distribution of the sample mean and variance, after some normalization. When the data are normal, we know the answer exactly, of course, but what about otherwise? We won’t answer that question quite yet, but take a step by looking at the joint distribution of the sample means of the \( X_i \)'s and the \( X_i^2 \)'s. We will assume that \( \text{Var}[X_i^2] < \infty \). Then we take
\[
W_n = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} \left( \frac{X_i}{X_i^2} \right) - \left( \frac{\mu^2 + \sigma^2}{\mu^2 + \sigma^2} \right) \right).
\] (16.47)

Then the central limit theorem says that
\[
W_n \xrightarrow{D} N(0, \text{Cov}(X_i, X_i^2)).
\] (16.48)

Look at that covariance. We know \( \text{Var}[X_i] = \sigma^2 \). Also,
\[
\text{Var}[X_i^2] = E[X_i^4] - E[X_i^2]^2 = \kappa_4 - (\mu^2 + \sigma^2)^2,
\] (16.49)
and
\[
\text{Cov}[X_i, X_i^2] = E[X_i^3] - E[X_i]E[X_i^2] = \kappa_3 - \mu(\mu^2 + \sigma^2),
\] (16.50)
where for convenience we’ve defined

\[ \kappa_k = E[X_i^k]. \]  

(16.51)

It’s not pretty, but the final answer is

\[
\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} \left( \frac{X_i}{X_i^2} \right) - \left( \frac{\mu}{\mu^2 + \sigma^2} \right) \right) \rightarrow^D N(0, n\left( \kappa_3 - \mu(\mu^2 + \sigma^2) \kappa_4 - (\mu^2 + \sigma^2)^2 \right)).
\]  

(16.53)

16.4 Mapping

Lemma 24 and equation (15.31) show that if \( W_n \rightarrow^P c \) then \( g(W_n) \rightarrow^P g(c) \) if \( g \) is continuous at \( c \). Similar results hold for convergence in distribution, and for combinations of convergences.

**Lemma 26. Some mapping lemmas**

1. If \( W_n \rightarrow^D W \), and \( g: \mathbb{R} \rightarrow \mathbb{R} \) is continuous at all the points in \( W \), the space of \( W \), then

\[ g(W_n) \rightarrow^D g(W). \]  

(16.54)

2. Similarly, for multivariate \( W_n \) (\( p \times 1 \)) and \( g \) (\( q \times 1 \)): If \( W_n \rightarrow^D W \) and \( g: \mathbb{R}^p \rightarrow \mathbb{R}^q \) is continuous at all the points in \( W \), then

\[ g(W_n) \rightarrow^D g(W). \]  

(16.55)

3. The next results constitute what is usually called **Slutsky’s Theorem**, or sometimes **Cramér’s Theorem**.

Suppose that

\[ Z_n \rightarrow^P c \quad \text{and} \quad W_n \rightarrow^D W. \]  

(16.56)

Then

\[ Z_n + W_n \rightarrow^D c + W, \quad Z_n W_n \rightarrow^D cW, \quad \text{and, if} \quad c \neq 0, \quad \frac{W_n}{Z_n} \rightarrow^D \frac{W}{c}. \]  

(16.57)

4. Generalizing Slutsky, if \( Z_n \rightarrow^P c \) and \( W_n \rightarrow^D W \), and \( g: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \) is continuous at \( \{c\} \times W \), then

\[ g(Z_n, W_n) \rightarrow^D g(c, W). \]  

(16.58)

5. Finally, the multivariate version of #4. If \( Z_n \rightarrow^P c \) (\( p_1 \times 1 \)) and \( W_n \rightarrow^D W \) (\( p_2 \times 1 \)), and \( g: \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \rightarrow \mathbb{R}^q \) is continuous at \( \{c\} \times W \), then

\[ g(Z_n, W_n) \rightarrow^D g(c, W). \]  

(16.59)
All the other four follow from #2, because convergence in probability is the same as convergence in distribution to a constant random variable. Also, #2 is just the multivariate version of #1, so basically they are all the same. They appear various places in various forms, though. The idea is that as long as the function is continuous, the limit of the function is the function of the limit.

Together with the law of large numbers and central limit theorem, these mapping results can prove a huge number of useful approximations. The $t$-statistic is one example. Let $X_1, \ldots, X_n$ be iid with mean $\mu$ and variance $\sigma^2 \in (0, \infty)$. Student’s $t$ statistic is defined as

$$T_n = \sqrt{n} \frac{\overline{X}_n - \mu}{S_n}, \quad \text{where} \quad S_n^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n - 1}. \quad (16.60)$$

We know that if the data are normal, this $T_n \sim t_{n-1}$ exactly, but if the data are not normal, who knows what the distribution is. We can find the limit, though, using Slutsky. Take

$$Z_n = S_n \quad \text{and} \quad W_n = \sqrt{n} \left( \overline{X}_n - \mu \right). \quad (16.61)$$

Then

$$Z_n \convP \sigma \quad \text{and} \quad W_n \convD N(0, \sigma^2) \quad (16.62)$$

from the homework and the central limit theorem, respectively. Because $\sigma^2 > 0$, the final component of (16.57) shows that

$$T_n = \frac{W_n}{Z_n} \convD \frac{N(0, \sigma^2)}{\sigma} = N(0, 1). \quad (16.63)$$

Thus for large $n$, $T_n$ is approximately $N(0, 1)$ even if the data are not normal. Thus

$$\overline{X}_n \pm 2 \frac{S_n}{\sqrt{n}} \quad (16.64)$$

is an approximate 95% confidence interval for $\mu$. Notice that this result doesn’t say anything about small $n$, especially it doesn’t say that the $t$ is better than the $z$ when the data are not normal. Other studies have shown that the $t$ is fairly robust, so it can be used at least when the data are approximately normal. Actually, heavy tails for the $X_i$’s means light tails for the $T_n$, so the $z$ might be better than $t$ in that case.

Regression through the origin.

Recall the example in (15.36) and (15.38). We know $\hat{\beta}_n$ is a consistent estimator of $\beta$, but what about its asymptotic distribution? That is,

$$\sqrt{n} \left( \hat{\beta}_n - \beta \right) \convD ?? \quad (16.65)$$
We need to do some manipulation to get it into a form where we can use the central limit theorem, etc. To that end,

\[ \sqrt{n} (\hat{\beta}_n - \beta) = \sqrt{n} \left( \frac{\sum_{i=1}^{n} X_i Y_i - \beta}{\sum_{i=1}^{n} X_i^2} \right) \]

\[ = \sqrt{n} \left( \frac{\sum_{i=1}^{n} X_i Y_i - \beta \sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^2} \right) \]

\[ = \frac{\sqrt{n} \sum_{i=1}^{n} (X_i Y_i - \beta X_i^2)}{\sum_{i=1}^{n} X_i^2 / n} \]  

(16.66)

The numerator in the last expression contains the sample mean of the \((X_i Y_i - \beta X_i^2)\)'s. Conditionally, from (15.38)

\[ E[X_i Y_i - \beta X_i^2 | X_i = x_i] = x_i \beta - \beta x_i^2 = 0 \]

so that unconditionally,

\[ E[X_i Y_i - \beta X_i^2] = 0, \ Var[X_i Y_i - \beta X_i^2] = E[X_i^2 \sigma^2] + Var[0] = \sigma^2 (\sigma_X^2 + \mu_X^2). \]  

(16.67)

Thus the central limit theorem shows that

\[ \sqrt{n} \sum_{i=1}^{n} (X_i Y_i - \beta X_i^2) / n \longrightarrow^D N(0, \sigma^2 (\sigma_X^2 + \mu_X^2)). \]  

(16.69)

We already know from (15.41) that \(\sum X_i^2 / n \longrightarrow^D \sigma_X^2 + \mu_X^2\), hence by Slutsky (16.57),

\[ \sqrt{n} (\hat{\beta}_n - \beta) \longrightarrow^D N \left( 0, \frac{\sigma^2 (\sigma_X^2 + \mu_X^2)}{\sigma_X^2 + \mu_X^2} \right). \]  

(16.70)

16.5 The \(\Delta\)-method

Recall the use of the \(\Delta\)-method to obtain the approximate variance of a function of a random variable, as in (6.106). Here we give a formal version of it.

**Lemma 27. \(\Delta\)-method.** Suppose

\[ \sqrt{n} (X_n - \mu) \longrightarrow^D W, \]  

(16.71)

and the function \(g : \mathbb{R} \rightarrow \mathbb{R}\) has a continuous derivative at \(\mu\). Then

\[ \sqrt{n} (g(X_n) - g(\mu)) \longrightarrow^D g'(\mu) W. \]  

(16.72)

**Proof.** As in (6.103), we have that

\[ g(X_n) = g(\mu) + (X_n - \mu) g'(\mu_n^*), \quad \mu_n^* \text{ is between } X_n \text{ and } \mu, \]  

(16.73)
hence
\[ \sqrt{n} \left( g(X_n) - g(\mu) \right) = \sqrt{n} \left( X_n - \mu \right) g'(\mu^*_n). \] (16.74)

We wish to show that \( g'(\mu^*_n) \to^p g'(\mu) \), but first need to show that \( X_n \to^p \mu \). Now
\[
X_n - \mu = \left[ \sqrt{n} \left( X_n - \mu \right) \right] \times \frac{1}{\sqrt{n}} \\
\to^D W \times 0 \quad \text{(because } \frac{1}{\sqrt{n}} \to^P 0) \\
= 0. \tag{16.75}
\]

That is, \( X_n - \mu \to^P 0 \), hence \( X_n \to^P \mu \). Because \( \mu^*_n \) is trapped between \( X_n \) and \( \mu \), \( \mu^*_n \to^P \mu \), which by continuity of \( g' \) means that
\[ g'(\mu^*_n) \to^P g'(\mu). \tag{16.76} \]

Applying Slutsky (16.57) to (16.74), by (16.71) and (16.76),
\[ \sqrt{n} \left( X_n - \mu \right) g'(\mu^*_n) \to^D W g'(\mu), \tag{16.77} \]
which is (16.72).

Usually, the limiting \( W \) is normal, so that under the conditions on \( g \), we have that
\[ \sqrt{n} \left( X_n - \mu \right) \to^D N(0, \sigma^2) \Rightarrow \sqrt{n} \left( g(X_n) - g(\mu) \right) \to^D N(0, g'(\mu)^2 \sigma^2). \tag{16.78} \]

### 16.5.1 The median

In Section 6.3.1, we used the \( \Delta \)-method to estimate the variance of order statistics. Here we apply Lemma 27 to the sample median. We have \( X_1, \ldots, X_n \) iid with distribution function \( F \). Let \( \eta \) be the median, so that \( F(\eta) = \frac{1}{2} \), and assume that the density \( f \) is positive and continuous at \( \eta \). For simplicity we take \( n \) odd, \( n = 2k - 1 \) for \( k \) a positive integer, so that \( X_{(k)} \), the \( k \)th order statistic, is the median. From (6.100), \( X_{(k)} \) has the same distribution as \( F^{-1}(U_{(k)}) \), where \( U_{(k)} \) is the median of a sample of \( n \) iid Uniform(0,1)'s. Also, from (6.97),
\[ U_{(k)} \sim Beta\left( \frac{n+1}{2}, \frac{n+1}{2} \right) = Beta(k, k). \tag{16.79} \]

The objective is to find the asymptotic distribution of \( X_{(k)} \), or actually
\[ \sqrt{n} \left( X_{(k)} - \eta \right) \text{ as } n \to \infty \text{ (i.e., } k \to \infty). \tag{16.80} \]

We start with the asymptotic distribution of \( U_{(k)} \). It can be shown that
\[ \sqrt{k} \left( U_{(k)} - \frac{1}{2} \right) \to N(0, \frac{1}{8}). \tag{16.81} \]

Thus for a function \( g \), with continuous derivative at \( \frac{1}{2} \), the \( \Delta \)-method (16.78) shows that
\[ \sqrt{k} \left( g(U_{(k)}) - g\left( \frac{1}{2} \right) \right) \to N(0, 8g'(\frac{1}{2})^2). \tag{16.82} \]
Let \( g(u) = F^{-1}(u) \). Then
\[ g(U_{(k)}) = F^{-1}(U_{(k)}) = D X_{(k)} \quad \text{and} \quad g\left(\frac{1}{2}\right) = F^{-1}\left(\frac{1}{2}\right) = \eta. \quad (16.83) \]

We also use the fact that
\[ g'(u) = \frac{1}{F'(F^{-1}(u))} \Rightarrow g'(\frac{1}{2}) = \frac{1}{f(\eta)}. \quad (16.84) \]

Thus making the substitutions in (16.82), and recalling that \( n = 2k - 1 \), we obtain for (16.80) that
\[
\sqrt{n} \left( X_{(k)} - \eta \right) = \sqrt{\frac{2k - 1}{k}} \sqrt{k} \left( X_{(k)} - \eta \right) \\
\rightarrow \sqrt{2} N(0, \frac{1}{8f(\eta)^2}),
\]

hence
\[
\sqrt{n} \left( X_{(k)} - \eta \right) \rightarrow N(0, \frac{1}{4f(\eta)^2}). \quad (16.86)
\]
(Compare the asymptotic variance here to the approximation in (6.99).)

**16.5.2 Variance stabilizing transformations**

Often, the variance of an estimator depends on the value of the parameter begin estimated. For example, if \( X_n \sim \text{Binomial}(n, p) \), then with \( \hat{p}_n = X_n/n \),
\[
\text{Var} \left[ \hat{p}_n \right] = \frac{p(1-p)}{n}. \quad (16.87)
\]

In regression situations, for instance, one usually desires the dependent \( Y_i \)'s to have the same variance for each \( i \), so that if these \( Y_i \)'s are binomial, or Poisson, the variance will not be constant. However, taking a function of the variables may achieve approximately equal variances. Such a function is called a variance stabilizing transformation. Formally, if \( \hat{\theta}_n \) is an estimator of \( \theta \), then we wish to find a \( g \) such that
\[
\sqrt{n} \left( g(\hat{\theta}_n) - g(\theta) \right) \longrightarrow^D N(0, 1). \quad (16.88)
\]

The “1” for the variance is arbitrary. The important thing is that it does not depend on \( \theta \).

In the binomial example, the variance stabilizing \( g \) would satisfy
\[
\sqrt{n} \left( g(\hat{p}_n) - g(p) \right) \longrightarrow^D N(0, 1). \quad (16.89)
\]

We know that
\[
\sqrt{n} \left( \hat{p}_n - p \right) \longrightarrow^D N(0, p(1-p)), \quad (16.90)
\]

and by the \( \Delta \)-method,
\[
\sqrt{n} \left( g(\hat{p}_n) - g(p) \right) \longrightarrow^D N(0, g'(p)^2 p(1-p)). \quad (16.91)
\]
What should \( g \) be so that that variance is 1? We need to solve
\[
g'(p) = \frac{1}{\sqrt{p(1-p)}},
\]
so that
\[
g(p) = \int_0^p \frac{1}{\sqrt{y(1-y)}}
\]
First, let \( u = \sqrt{y} \), so that \( y = u^2 \) and \( dy = 2udu \), and
\[
g(p) = \int_0^{\sqrt{p}} \frac{1}{u \sqrt{1-u^2}}
\]
The integral is \( \arcsin(u) \), which means the variance stabilizing transformation is
\[
g(p) = 2 \arcsin(\sqrt{p}).
\]
Note that adding a constant to \( g \) won’t change the derivative. The approximation suggested by (16.89) is then
\[
2 \arcsin(\sqrt{\hat{p}_n}) \approx N(2 \arcsin(\sqrt{p}), \frac{1}{n}).
\]
An approximate 95% confidence interval for \( 2 \arcsin(\sqrt{p}) \) is
\[
2 \arcsin(\sqrt{\hat{p}_n}) \pm \frac{2}{\sqrt{n}}.
\]
That interval can be inverted to obtain the interval for \( p \), that is, apply \( g^{-1}(u) = \sin(u/2)^2 \) to both ends:
\[
p \in \left( \sin(\arcsin(\sqrt{\hat{p}_n}) - \frac{1}{\sqrt{n}})^2, \sin(\arcsin(\sqrt{\hat{p}_n}) + \frac{1}{\sqrt{n}})^2 \right).
\]
This interval may be slightly better than the usual approximate interval,
\[
\hat{p}_n \pm 2 \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}.
\]

### 16.6 The MLE in Exponential Families

Consider a regular one-dimensional exponential family model, where \( \theta \) is the natural parameter and \( x \) is the natural statistic, i.e., \( X_1, \ldots, X_n \) are iid with pdf
\[
f(x \mid \theta) = a(x)e^{\theta x-\psi(\theta)}.
\]
We know that the maximum likelihood estimator of \( \theta \) is
\[
\hat{\theta}_n = \mu^{-1}(\bar{X}_n), \text{ where } \mu(\theta) = E_{\theta}[X_1] = \psi'(\theta).
\]
Also, by the central limit theorem,
\[
\sqrt{n} (\bar{X}_n - \mu(\theta)) \overset{D}{\longrightarrow} N(0, I_1(\theta)),
\]
where $I_1$ is the Fisher information in one observation:

$$I_1(\theta) = \mu'(\theta) = \psi''(\theta) = \text{Var}_\theta[X_i]. \quad (16.103)$$

The asymptotic distribution of $\hat{\theta}_n$ can then be found using the $\Delta$-method with $g(w) = \mu^{-1}(w)$. As in (6.107),

$$g'(w) = \frac{1}{\mu'\mu^{-1}(w)}, \quad (16.104)$$

(16.102) implies

$$\sqrt{n} \left( \mu^{-1}(\overline{X}_n) - \mu^{-1}(\mu(\theta)) \right) \xrightarrow{D} N(0, \frac{1}{\mu'(\mu^{-1}(\mu(\theta)))^2 I_1(\theta)}). \quad (16.105)$$

But,

$$\hat{\theta}_n = \mu^{-1}(\overline{X}_n), \quad \mu^{-1}(\mu(\theta)) = \theta, \quad (16.106)$$

and

$$\frac{1}{\mu'\mu^{-1}(\mu(\theta)))^2} I_1(\theta) = \frac{1}{\mu'(\theta)^2} I_1(\theta) = \frac{1}{I_1(\theta)} I_1(\theta) = \frac{1}{I_1(\theta)} \quad (16.107)$$

That is,

$$\sqrt{n} (\hat{\theta}_n - \theta) \xrightarrow{D} N(0, \frac{1}{I_1(\theta)}), \quad (16.108)$$

so that for large $n$,

$$\hat{\theta}_n \approx N(\theta, \frac{1}{n I_1(\theta)}). \quad (16.109)$$

Notice that, at least asymptotically, the MLE achieves the Cramér-Rao lower bound. Not that the MLE is unbiased, but it at least is consistent.

One limitation of the result (16.109) is that the asymptotic variance depends on the unknown parameter. But we do can consistently estimate the parameter, and $I_1(\theta)$ is a continuous function, hence

$$I_1(\hat{\theta}_n) \xrightarrow{D} I_1(\theta), \quad (16.110)$$

so that Slutsky can be used to show that from (16.108),

$$\sqrt{n \frac{1}{I_1(\hat{\theta}_n)}} (\hat{\theta}_n - \theta) \xrightarrow{D} N(0, 1), \quad (16.111)$$

and an approximate 95% confidence interval for $\theta$ is

$$\hat{\theta}_n \pm \frac{2}{\sqrt{n I_1(\hat{\theta}_n)}}. \quad (16.112)$$

What about functions $g(\theta)$? We know the MLE of $g(\theta)$ is $g(\hat{\theta}_n)$, and if $g'$ is continuous,

$$\sqrt{n} (g(\hat{\theta}_n) - g(\theta)) \xrightarrow{D} N(0, \frac{g'(\theta)^2}{I_1(\theta)}). \quad (16.113)$$

Notice that that variance is also the CRLB for unbiased estimates of $g(\theta)$, at least if you divide by $n$. 
16.7 The multivariate \( \Delta \)-method

It might be that we have a function of several random variables to deal with, or more generally we have several functions of several variables. For example, we might be interested in the mean and variance simultaneously, as in (15.4). So we start with a sequence of \( p \times 1 \) random vector, whose asymptotic distribution is multivariate normal:

\[
\sqrt{n} \left( X_n - \mu \right) \longrightarrow^D N(0, \Sigma), \tag{16.114}
\]

and a function \( g : \mathbb{R}^p \rightarrow \mathbb{R}^q \). We cannot just take the derivative of \( g \), since there are \( pq \) of them. What we need is the entire matrix of derivatives, just as for finding the Jacobian. Letting

\[
g(w) = \begin{pmatrix}
g_1(w) \\
g_2(w) \\
\vdots \\
g_q(w)
\end{pmatrix}, \quad w = \begin{pmatrix}
w_1 \\
w_2 \\
\vdots \\
w_p
\end{pmatrix}, \tag{16.115}
\]

define the \( p \times q \) matrix

\[
D(w) = \begin{pmatrix}
\frac{\partial}{\partial w_1} g_1(w) & \frac{\partial}{\partial w_2} g_1(w) & \cdots & \frac{\partial}{\partial w_q} g_1(w) \\
\frac{\partial}{\partial w_1} g_2(w) & \frac{\partial}{\partial w_2} g_2(w) & \cdots & \frac{\partial}{\partial w_q} g_2(w) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial}{\partial w_1} g_q(w) & \frac{\partial}{\partial w_2} g_q(w) & \cdots & \frac{\partial}{\partial w_q} g_q(w)
\end{pmatrix}. \tag{16.116}
\]

**Lemma 28. Multivariate \( \Delta \)-method.** Suppose (16.114) holds, and \( D(w) \) in (16.116) is continuous at \( w = \mu \). Then

\[
\sqrt{n} \left( g(X_n) - g(\mu) \right) \longrightarrow^D N(0, D(\mu)^\top \Sigma D(\mu)). \tag{16.117}
\]

The \( \Sigma \) is \( p \times p \) and \( D \) is \( p \times q \), so that the covariance in (16.117) is \( q \times q \), as it should be. Some examples follow.

**Mean and variance**

Go back to the example that ended with (16.53): \( X_1, \ldots, X_n \) are iid with mean \( \mu \), variance \( \sigma^2 \), \( E[X_i^3] = \kappa_3 \) and \( E[X_i^4] = \kappa_4 \). Ultimately, we wish the asymptotic distribution of \( \overline{X}_n \) and \( S_n^2 \), so we start with that of \( (\sum X_i, \sum X_i^2) \), so that

\[
X_n = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} X_i \\ X_i^2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu \\ \mu^2 + \sigma^2 \end{pmatrix} \tag{16.118}
\]

and

\[
\Sigma = \begin{pmatrix} \sigma^2 & \kappa_3 - \mu(\mu^2 + \sigma^2) \\ \kappa_3 - \mu(\mu^2 + \sigma^2) & \kappa_4 - (\mu^2 + \sigma^2)^2 \end{pmatrix}. \tag{16.119}
\]

With \( S_n^2 = \sum X_i^2 / n - \overline{X}_n^2 \),

\[
\begin{pmatrix} \overline{X}_n \\ S_n^2 \end{pmatrix} = \begin{pmatrix} g_1(X_n, \sum X_i^2 / n) \\ g_2(X_n, \sum X_i^2 / n) \end{pmatrix}, \tag{16.120}
\]
where
\[ g_1(w_1, w_2) = w_1 \quad \text{and} \quad g_2(w_1, w_2) = w_2 - w_1^2. \quad (16.121) \]

Then
\[ D(\mathbf{w}) = \begin{pmatrix} \frac{\partial w_1}{\partial w_1} & \frac{\partial (w_2 - w_1^2)}{\partial w_1} \\ \frac{\partial w_1}{\partial w_2} & \frac{\partial (w_2 - w_1^2)}{\partial w_2} \end{pmatrix} = \begin{pmatrix} 1 & -2w_1 \\ 0 & 1 \end{pmatrix}. \quad (16.122) \]

Also,
\[ g(\mu) = \begin{pmatrix} \sigma^2 + \mu^2 - \mu^2 \end{pmatrix} = \begin{pmatrix} \mu \sigma^2 \end{pmatrix} \quad \text{and} \quad D(\mu) = \begin{pmatrix} 1 & -2\mu \end{pmatrix}, \quad (16.123) \]

and
\[ D(\mu) \Sigma D(\mu) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}. \quad (16.125) \]

hence
\[ \sqrt{n} \left( \begin{pmatrix} \overline{X}_n \\ S_n^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \right) \rightarrow^D N(0_2, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}). \quad (16.126) \]

Actually, that is not surprising, since we know the variance of \( \overline{X}_n \) is \( \sigma^2 / n \), and that of \( S_n^2 \) is the variance of a \( \chi^2_{n-1} \) times \( (n-1)\sigma^2 / n \), which is \( 2(n-1)^2 \sigma^4 / n^2 \). Multiplying those by \( n \) and letting \( n \to \infty \) yields the diagonals \( \sigma^2 \) and \( 2\sigma^4 \). Also the mean and variance are independent, so their covariance is 0.

From these, we can find the \textit{coefficient of variance}, or the noise-to-signal ratio,
\[ cv = \frac{\sigma}{\mu} \quad \text{and \ sample \ version} \quad \hat{cv} = \frac{S_n}{\overline{X}_n}. \quad (16.127) \]

Then \( \hat{cv} = h(\overline{X}_n, S_n^2) \) where \( h(w_1, w_2) = \sqrt{w_2} / w_1 \). The derivatives here are
\[ D_h(\mathbf{w}) = \begin{pmatrix} -\sqrt{w_2} \\ -\frac{w_2}{2\sqrt{w_2}} \end{pmatrix} \quad \Rightarrow \quad D_h(\mu, \sigma^2) = \begin{pmatrix} -\frac{\sigma}{\mu^2} \\ -\frac{\sigma}{2\sigma\mu} \end{pmatrix}. \quad (16.128) \]

Then, assuming that \( \mu \neq 0 \),
\[ D_h^2 \Sigma D_h = \begin{pmatrix} \frac{\sigma^4}{\mu^4} & 0 \\ 0 & \frac{\sigma^4}{2\mu^2} \end{pmatrix}. \quad (16.129) \]
Chapter 16. Asymptotics: Convergence in distribution

The asymptotic distribution is then
\[
\sqrt{n} (\hat{c}v - cv) \rightarrow^D N(0, \frac{\sigma^4}{\mu^4} + \frac{1}{2} \frac{\sigma^2}{\mu^2}) = N(0, cv^2(cv^2 + \frac{1}{2})).
\] (16.130)

For example, data on \(n = 102\) female students’ heights had a mean of 65.56 and a standard deviation of 2.75, so the \(\hat{c}v = 2.75/65.56 = 0.0419\). We can find a confidence interval by estimating the variance in (16.130) in the obvious way:
\[
(\hat{c}v \pm 2 \left| \frac{\hat{c}v \sqrt{0.5 + \hat{c}v^2}}{\sqrt{n}} \right|) = (0.0419 \pm 2 \times 0.0029) = (0.0361, 0.0477),
\] (16.131)

For the men, the mean is 71.25 and the sd is 2.94, so their cv is 0.0413. That’s practically the same as for the women. The men’s standard error of \(\hat{c}v\) is 0.0037 (the \(n = 64\)), so a confidence interval for the difference between the women and men is
\[
(0.0419 - 0.0413 \pm 2 \sqrt{0.0029^2 + 0.0037^2}) = (0.0006 \pm 0.0094).
\] (16.132)

Clearly 0 is in that interval, so there does not appear to be any difference between the coefficients of variation.

Correlation coefficient

Consider the bivariate normal sample,
\[
\left( \begin{array}{c} X_1 \\ Y_1 \\ \vdots \\ X_n \end{array} \right), \ldots, \left( \begin{array}{c} X_n \\ Y_n \end{array} \right) \text{ are iid } \sim N(\mathbf{Q}_2, \left( \begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array} \right)).
\] (16.133)

The sample correlation coefficient in this case (we don’t need to subtract the means) is
\[
r_n = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i^2}}.
\] (16.134)

From the homework, we know that \(r_n \rightarrow^P \rho\). What about the asymptotic distribution? Notice that \(r_n\) can be written as a function of three sample means,
\[
r_n = g(\sum_{i=1}^{n} X_i Y_i / n, \sum_{i=1}^{n} X_i^2 / n, \sum_{i=1}^{n} Y_i^2 / n) \text{ where } g(w_1, w_2, w_3) = \frac{w_1}{\sqrt{w_1 w_2}}.
\] (16.135)

First, apply the central limit theorem to the three means:
\[
\sqrt{n} \left( \left( \begin{array}{c} \sum_{i=1}^{n} X_i Y_i / n \\ \sum_{i=1}^{n} X_i^2 / n \\ \sum_{i=1}^{n} Y_i^2 / n \end{array} \right) - \mu \right) \rightarrow^D N(\mathbf{0}_3, \Sigma).
\] (16.136)

Now \(E[X_i Y_i] = \rho\) and \(E[X_i^2] = E[Y_i^2] = 1\), hence
\[
\mu = \left( \begin{array}{c} \rho \\ 1 \\ 1 \end{array} \right).
\] (16.137)
16.7. The multivariate $\Delta$-method

The covariance is a little more involved. The homework shows $\text{Var}[X_i Y_i] = 1 + \rho^2$. Because $X_i$ and $Y_i$ are $N(0,1)$, their squares are $\chi_1^2$, whose variance is 2. Next,

$$
\text{Cov}[X_i Y_i, X_i^2] = E[X_i^3 Y_i] - E[X_i Y_i] E[X_i^2] = E[E[X_i^3 Y_i | X_i]] - \rho = E[X_i^3 \rho X_i] - \rho = 3\rho - \rho = 2\rho,
$$

(16.138)

because $E[X_i^4] = \text{Var}[X_i^4] + E[X_i^2]^2 = 2 + 1$. By symmetry, $\text{Cov}[X_i Y_i, Y_i^2] = 2\rho$, too. Finally,

$$
\text{Cov}[Y_i^2, X_i^2] = E[X_i^2 Y_i^2] - E[Y_i^2] E[X_i^2] = E[E[X_i^2 Y_i^2 | X_i]] - 1 = E[X_i^3 \rho X_i^2 + (1 - \rho^2)] - 1 = 2\rho^2.
$$

(16.139)

Putting those together we have that

$$
\Sigma = \text{Cov}
\begin{pmatrix}
X_i Y_i
X_i^2 Y_i
Y_i^2
\end{pmatrix}
= \begin{pmatrix}
1 + \rho^2 & 2\rho & 2\rho \\
2\rho & 2 & 2\rho^2 \\
2\rho & 2\rho^2 & 2
\end{pmatrix}.
$$

(16.140)

Now for the derivatives of $g(w) = w_2/\sqrt{w_2 w_3}$:

$$
D(w) = \begin{pmatrix}
1 \\
\frac{1}{\sqrt{w_2 w_3}} \\
\frac{-1}{2} \frac{w_3}{w_2 w_3} \\
\frac{-1}{2} \frac{w_2}{w_3} \sqrt{w_3}
\end{pmatrix} \implies D(\mu) = D(\rho, 1, 1) = \begin{pmatrix}
1 \\
-\frac{1}{2} \rho \\
-\frac{1}{2} \rho
\end{pmatrix}.
$$

(16.141)

The $\Delta$-method applied to (16.136) shows that

$$
\sqrt{n} (r_n - \rho) \longrightarrow^D N(0, (1 - \rho^2)^2),
$$

(16.142)

where that asymptotic variance is found by just writing out

$$
D(\mu)^T \Sigma D(\mu) = \begin{pmatrix}
1 & -\frac{1}{2} \rho & -\frac{1}{2} \rho
\end{pmatrix} \begin{pmatrix}
1 + \rho^2 & 2\rho & 2\rho \\
2\rho & 2 & 2\rho^2 \\
2\rho & 2\rho^2 & 2
\end{pmatrix} \begin{pmatrix}
-\frac{1}{2} \rho \\
-\frac{1}{2} \rho
\end{pmatrix} = (1 - \rho^2)^2.
$$

(16.143)

**Linear functions**

If $A$ is a $q \times p$ matrix, then it is easy to get the asymptotic distribution of $AX_n$:

$$
\sqrt{n} (X_n - \mu) \longrightarrow^D N(0_n, \Sigma) \implies \sqrt{n} (AX_n - A\mu) \longrightarrow^D N(0_n, A\Sigma A'),
$$

(16.144)

because for the function $g(w) = Aw$, $D(w) = A'$. 

Chapter 17

Likelihood Estimation

The challenge when trying to estimate a parameter is to first find an estimator, and second to find a good one, and third to find its standard error, and fourth to find its distribution. We have found many nice estimators, but faced with a new situation, what does one do? Fortunately, there is a solution to all four challenges, at least when \( n \) is large (and certain conditions hold): the maximum likelihood estimator. We have already seen this result in the regular one-dimensional exponential family case (Section 16.6), where

\[
\sqrt{n} (\hat{\theta}_n - \theta) \xrightarrow{D} N(0, \frac{1}{I_1(\theta)}). \tag{17.1}
\]

Plus we have that

\[
I_1(\hat{\theta}_n) \xrightarrow{P} I_1(\theta), \tag{17.2}
\]

hence

\[
\sqrt{nI_1(\hat{\theta}_n)} (\hat{\theta}_n - \theta) \xrightarrow{D} N(0,1). \tag{17.3}
\]

Thus the MLE is asymptotically normal, consistent, asymptotically efficient, and we can estimate the standard error effectively. What happens outside of exponential families?

17.1 The setup: Cramér’s conditions

We need a number of technical assumptions to hold. They easily hold in exponential families, but for other pdf’s they may or may not be easy to verify. We start with \( X_1, \ldots, X_n \) iid, each with space \( \mathcal{X}_0 \) and pdf \( f(x \mid \theta), \theta \in \Theta = (a,b) \). First, we need that the space of \( X_i \) is the same for each \( \theta \), which is satisfied if

\[
f(x \mid \theta) > 0 \text{ for all } x \in \mathcal{X}_0, \theta \in \Theta. \tag{17.4}
\]

This requirement eliminates the \( \text{Uniform}(0,\theta) \), for example. Which is not to say that the MLE is bad in this case, but that the asymptotic normality, etc., does not hold. We also need that

\[
\frac{\partial f(x \mid \theta)}{\partial \theta}, \frac{\partial^2 f(x \mid \theta)}{\partial \theta^2}, \frac{\partial^3 f(x \mid \theta)}{\partial \theta^3} \text{ exist for all } x \in \mathcal{X}_0, \theta \in \Theta. \tag{17.5}
\]
In order for the score and information functions to exist and behave correctly, assume that for any \( \theta \in \Theta \),
\[
\int_{X} \frac{\partial f(x \mid \theta)}{\partial \theta} dx = \frac{\partial}{\partial \theta} \int_{X} f(x \mid \theta) dx = 0 \quad (17.6)
\]
and
\[
\int_{X} \frac{\partial^2 f(x \mid \theta)}{\partial^2 \theta} dx = \frac{\partial^2}{\partial^2 \theta} \int_{X} f(x \mid \theta) dx = 0. \quad (17.7)
\]
(Replace the integrals with sums for the discrete case.) Denote the loglikelihood for one observation by \( l \):
\[
l(\theta; x) = \log(f(x \mid \theta)). \quad (17.8)
\]
Define the score function by
\[
S(\theta; x) = l'(\theta; x), \quad (17.9)
\]
and the Fisher information in one observation by
\[
I_1(\theta) = E_{\theta}[S(\theta; X)^2]. \quad (17.10)
\]
Assume that
\[
0 < I_1(\theta) < \infty \quad \text{for all } \theta \in \Theta. \quad (17.11)
\]
We have the following, which was used in deriving the Cramér-Rao lower bound.

**Lemma 29.** If (17.4), (17.6), (17.7), and (17.10) hold, then
\[
E_{\theta}[S(\theta; X)] = 0, \quad \text{hence } \text{Var}_{\theta}[S(\theta, X)] = I_1(\theta), \quad (17.12)
\]
and
\[
I_1(\theta) = -E_{\theta}[l''(\theta; X)] = E_{\theta}[S'(\theta; X)]. \quad (17.13)
\]

**Proof.** First, by definition of the score function (17.9),
\[
E_{\theta}[S(\theta; X)] = E_{\theta} \left[ \frac{\partial}{\partial \theta} \log(f(x \mid \theta)) \right] = \int_{X} \frac{\partial f(x \mid \theta)/\partial \theta}{f(x \mid \theta)} f(x \mid \theta) dx = \int_{X} \frac{\partial f(x \mid \theta)}{\partial \theta} dx = 0 \quad \text{by (17.6)} \quad (17.14)
\]
Next, write
\[
E_{\theta}[l''(\theta; X)] = E_{\theta} \left[ \frac{\partial^2}{\partial \theta^2} \log(f(X|\theta)) \right] = E_{\theta} \left[ \frac{\partial^2 f(X|\theta)/\partial \theta^2}{f(X|\theta)} - \left( \frac{\partial f(X|\theta)/\partial \theta}{f(X|\theta)} \right)^2 \right] = \int_{X_0} \frac{\partial^2 f(x|\theta)/\partial^2 \theta}{f(x|\theta)} f(x|\theta) dx - E_{\theta}[S(\theta; X)]^2 = \frac{\partial^2}{\partial \theta^2} \int_{X_0} f(x|\theta) dx - E_{\theta}[S(\theta; X)]^2 = 0 - I_1(\theta), \quad (17.15)
\]
which proves (17.13).

One more technical assumption we need is that for each \( \theta \in \Theta \) (which will take the role of the “true” value of the parameter), there exists an \( \epsilon > 0 \) and a function \( M(x) \) such that

\[
|l'''(t; x)| \leq M(x) \text{ for } \theta - \epsilon < t < \theta + \epsilon, \quad \text{and} \quad E_{\theta}[M(X)] < \infty. \tag{17.16}
\]

### 17.2 Consistency

First we address the question of whether the MLE is a consistent estimator of \( \theta \). The short answer is “Yes,” although things can get sticky if there are multiple maxima. But before we get to the results, there are some mathematical prerequisites to deal with.

#### 17.2.1 Convexity and Jensen’s inequality

**Definition 30. Convexity.** A function \( g : \mathcal{X}_0 \to \mathbb{R} \), \( \mathcal{X}_0 \subset \mathbb{R} \), is convex if for each \( x_0 \in \mathcal{X} \), there exist \( a_0 \) and \( \beta_0 \) such that

\[
g(x_0) = a_0 + \beta_0 x_0, \tag{17.17}
\]

and

\[
g(x) \geq a_0 + \beta_0 x \text{ for all } x \in \mathcal{X}_0. \tag{17.18}
\]

The function is strictly convex if (17.17) holds, and

\[
g(x) > a_0 + \beta_0 x \text{ for all } x \in \mathcal{X}_0, \ x \neq x_0. \tag{17.19}
\]

The definition basically means that the tangent to \( g \) at any point lies below the function. If \( g''(x) \) exists for all \( x \), then \( g \) is convex if and only if \( g''(x) \geq 0 \) for all \( x \), and it is strictly convex if and only if \( g''(x) > 0 \) for all \( x \). Notice that the line need not be unique. For example, \( g(x) = |x| \) is convex, but when \( x_0 = 0 \), any line through \((0, 0)\) with slope between \( \pm 1 \) will lie below \( g \).

By the same token, any line segment connecting two points on the curve lies above the curve, as in the next lemma.

**Lemma 30.** If \( g \) is convex, \( x, y \in \mathcal{X}_0 \), and \( 0 < \epsilon < 1 \), then

\[
\epsilon g(x) + (1 - \epsilon)g(y) \geq g(\epsilon x + (1 - \epsilon)y). \tag{17.20}
\]

If \( g \) is strictly convex, then

\[
\epsilon g(x) + (1 - \epsilon)g(y) > g(\epsilon x + (1 - \epsilon)y) \text{ for } x \neq y. \tag{17.21}
\]

Rather than prove this lemma, we will prove the more general result for random variables.

**Lemma 31. Jensen’s inequality.** Suppose that \( X \) is a random variable with space \( \mathcal{X}_0 \), and that \( E[X] \) exists. If the function \( g \) is convex, then

\[
E[g(X)] \geq g(E[X]), \tag{17.22}
\]

where \( E[g(X)] \) may be \( +\infty \). Furthermore, if \( g \) is strictly convex and \( X \) is not constant,

\[
E[g(X)] > g(E[X]). \tag{17.23}
\]
Chapter 17. Likelihood Estimation

Proof. We’ll prove it just in the strictly convex case, when $X$ is not constant. The other case is easier. Apply Definition 30 with $x_0 = E[X]$, so that

$$g(E[X]) = a_0 + \beta_0 E[X], \quad \text{and} \quad g(x) > a_0 + \beta_0 x \quad \text{for all} \quad x \neq E[X].$$

(17.24)

But then

$$E[g(X)] > E[a_0 + \beta_0 X] = a_0 + \beta_0 E[X] = g(E[X]),$$

(17.25)

because there is a positive probability $X \neq E[X]$.

Why does Lemma 31 imply Lemma 30?1

A mnemonic device for which way the inequality goes is to think of the convex function $g(x) = x^2$. Jensen implies that

$$E[X^2] \geq E[X]^2,$$

(17.26)

but that is the same as saying that $Var[X] \geq 0$. Also, $Var[X] > 0$ unless $X$ is a constant.

Convexity is also defined for $x$ being a $p \times 1$ vector, so that $X_0 \subset \mathbb{R}^p$, in which case the line $a_0 + \beta_0 x$ in Definition 30 becomes a hyperplane $\alpha_0 + \beta^T x$. Jensen’s inequality follows as well, where we just turn $X$ into a vector.

17.2.2 A consistent sequence of roots

The loglikelihood function for the data is

$$l_n(\theta; x_1, \ldots, x_n) = \sum_{i=1}^n \log(f(x_i \mid \theta)) = \sum_{i=1}^n l(\theta; x_i).$$

(17.27)

For now, we assume that the space does not depend on $\theta$ (17.4), and the first derivative in (17.5) is continuous. Also, for each $n$ and $x_1, \ldots, x_n$, there exists a unique solution to $l_n'(\theta; x_1, \ldots, x_n) = 0$:

$$l_n'(\hat{\theta}_n; x_1, \ldots, x_n) = 0, \quad \hat{\theta}_n \in \Theta.$$

(17.28)

Note that this $\hat{\theta}_n$ is a function of $x_1, \ldots, x_n$. It is also generally the maximum likelihood estimate, although it is possible it is a local minimum or an inflection point rather than the maximum.

Now suppose $\theta$ is the true parameter, and take $\epsilon > 0$. Look at the difference, divided by $n$, of the likelihoods at $\theta$ and $\theta + \epsilon$:

$$\frac{1}{n} \left( l_n(\theta_0; x_1, \ldots, x_n) - l_n(\theta_0 + \epsilon; x_1, \ldots, x_n) \right) = \frac{1}{n} \sum_{i=1}^n \log \left( \frac{f(x_i \mid \theta)}{f(x_i \mid \theta + \epsilon)} \right) = \frac{1}{n} \sum_{i=1}^n - \log \left( \frac{f(x_i \mid \theta + \epsilon)}{f(x_i \mid \theta)} \right).$$

(17.29)

The final expression is the mean of iid random variables, hence by the WLLN it converges in probability to the expected value of the summand:

$$\frac{1}{n} \left( l_n(\theta; x_1, \ldots, x_n) - l_n(\theta + \epsilon; x_1, \ldots, x_n) \right) \rightarrow^P E_{\theta}[-\log \left( \frac{f(X \mid \theta + \epsilon)}{f(X \mid \theta)} \right)].$$

(17.30)

1Take $X$ to be the random variable with $P[X = x] = \epsilon$ and $P[X = y] = 1 - \epsilon$. 


17.2. Consistency

Now apply Jensen’s inequality, Lemma 31, to that expected value, with \( g(x) = -\log(x) \), and the random variable being \( f(X | \theta) / f(X | \theta + \epsilon) \). This \( g \) is strictly convex, and the random variable is not constant because the parameters are different, hence

\[
E_{\theta}[-\log \left( \frac{f(X | \theta + \epsilon)}{f(X | \theta)} \right)] > -\log \left( E_{\theta} \left[ \frac{f(X | \theta + \epsilon)}{f(X | \theta)} \right] \right)
\]

\[
= -\log \left( \int_{X_0} \frac{f(x | \theta + \epsilon)}{f(x | \theta)} f(x | \theta) dx \right)
\]

\[
= -\log \left( \int_{X_0} f(x | \theta + \epsilon) dx \right)
\]

\[
= -\log(1) = 0 \quad (17.31)
\]

The same result holds for \( \theta - \epsilon \), hence

\[
\frac{1}{n} \left( l_n(\theta; x_1, \ldots, x_n) - l_n(\theta + \epsilon; x_1, \ldots, x_n) \right) \to^P a > 0 \quad \text{and}
\]

\[
\frac{1}{n} \left( l_n(\theta; x_1, \ldots, x_n) - l_n(\theta - \epsilon; x_1, \ldots, x_n) \right) \to^P b > 0. \quad (17.32)
\]

These equations mean that eventually, the likelihood at \( \theta \) is higher than that at \( \theta \pm \epsilon \); precisely, dropping the \( X_i \)'s for convenience:

\[
P_{\theta} \left[ l_n(\theta) > l_n(\theta + \epsilon) \text{ and } l_n(\theta) > l_n(\theta - \epsilon) \right] \to 1. \quad (17.33)
\]

Note that if \( l_n(\theta) > l_n(\theta + \epsilon) \) and \( l_n(\theta) > l_n(\theta - \epsilon) \), then between \( \theta - \epsilon \) and \( \theta + \epsilon \), the likelihood goes up then comes down again. Because the derivative is continuous, somewhere between \( \theta \pm \epsilon \) the derivative must be 0. By assumption, that point is the unique root \( \hat{\theta}_n \). It is also the maximum. Which means that

\[
l_n(\theta) > l_n(\theta + \epsilon) \quad \text{and} \quad l_n(\theta) > l_n(\theta - \epsilon) \Rightarrow \theta - \epsilon < \hat{\theta}_n < \theta + \epsilon. \quad (17.34)
\]

By (17.33), the left hand side of (17.34) has probability going to 1, hence

\[
P[|\hat{\theta}_n - \theta| < \epsilon] \to 1 \Rightarrow \hat{\theta}_n \to^P \theta, \quad (17.35)
\]

and the MLE is consistent.

The requirement that there is a unique root (17.28) for all \( n \) and set of \( x_i \)'s is too strong. The main problem is that sometimes the maximum of the likelihood does not exist over \( \Theta = (a, b) \), but at \( a \) or \( b \). For example, in the Binomial case, if the number of successes is 0, then the MLE of \( p \) would be 0, which is not in (0,1). Thus in the next theorem, we need only that probably there is a unique root.

**Theorem 11.** Suppose that

\[
P_{\theta} \left[ l_n'(t; X_1, \ldots, X_n) \text{ has a unique root } \hat{\theta}_n \in \Theta \right] \to 1. \quad (17.36)
\]

Then

\[
\hat{\theta}_n \to^P \theta. \quad (17.37)
\]
Technically, if there is not a unique root, you can choose \( \hat{\theta}_n \) to be whatever you want, but typically it would be either one of a number of roots, or one of the limiting values \( a \) and \( b \). Equation (17.36) does not always hold. For example, in the Cauchy location-family case, the number of roots goes in distribution to \( 1 + \text{Poisson}(1/\pi) \) (ref\(^2\)), so there is always a good chance of two or more roots. But it will be true that if you pick the right root, it will be consistent.

### 17.3 Asymptotic normality

To find the asymptotic distribution of the MLE, we first expand the derivative of the likelihood around \( \theta = \hat{\theta}_n \):

\[
l_n'(\hat{\theta}_n) = l_n'(\theta) + (\hat{\theta}_n - \theta) l_n''(\theta) + \frac{(\hat{\theta}_n - \theta)^2}{2} l_n'''(\theta^*_n), \quad \theta^*_n \text{ between } \theta \text{ and } \hat{\theta}_n. \tag{17.38}
\]

If \( \hat{\theta}_n \) is a root of \( l_n' \) as in (17.28), then

\[
0 = l_n'(\theta) + (\hat{\theta}_n - \theta) l_n''(\theta) + \frac{(\hat{\theta}_n - \theta)^2}{2} l_n'''(\theta^*_n)
\]

\[
\implies (\hat{\theta}_n - \theta)(l_n''(\theta) + \frac{(\hat{\theta}_n - \theta)}{2} l_n'''(\theta^*_n)) = -l_n'(\theta)
\]

\[
\implies \sqrt{n} (\hat{\theta}_n - \theta) = -\frac{\sqrt{n} \frac{1}{n} l_n'(\theta)}{\frac{1}{n} l_n''(\theta) + (\hat{\theta}_n - \theta) \frac{1}{n} l_n'''(\theta^*_n)/2}. \tag{17.39}
\]

The task is then to find the limits of the three terms on the right: the numerator and the two summands in the denominator.

**Theorem 12. Asymptotic normality of the MLE (Cramér).** Suppose that the assumptions in Section 17.1 hold, i.e., (17.4), (17.5), (17.6), (17.7), (17.11), and (17.16). Also, suppose that \( \hat{\theta}_n \) is a consistent sequence of roots of (17.28), that is, \( l_n'(\hat{\theta}_n) = 0 \) and \( \hat{\theta}_n \rightarrow^D \theta \), where \( \theta \) is the true parameter. Then

\[
\sqrt{n} (\hat{\theta}_n - \theta) \longrightarrow^D N(0, \frac{1}{I_1(\theta)}). \tag{17.40}
\]

**Proof.** First,

\[
\frac{1}{n} l'(\theta) = \frac{1}{n} \sum_{i=1}^{n} l'(\theta; x_i) = \frac{1}{n} \sum_{i=1}^{n} S(\theta; x_i). \tag{17.41}
\]

The scores \( S(\theta; X_i) \) are iid with mean 0 and variance \( I_1(\theta) \) as in (17.12), so that the central limit theorem implies that

\[
\sqrt{n} \frac{1}{n} l'(\theta) \longrightarrow^D N(0, I_1(\theta)). \tag{17.42}
\]

Second,

\[
\frac{1}{n} l''_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} l''_n(\theta; x_i). \tag{17.43}
\]

\(^2\text{Reeds (1985), Annals of Statistics, pages 775–784.}\)
The $l''(\theta; X_i)$’s are iid with mean $-I_1(\theta)$ by (17.13), hence the WLLN shows that

$$\frac{1}{n} l''_n(\theta) \to^P -I_1(\theta).$$  \hfill (17.44)

**Third,** consider the $M(x_i)$ from assumption (17.16). By the WLLN,

$$\frac{1}{n} \sum_{i=1}^{n} M(X_i) \to^P E_\theta[M(X)] < \infty,$$  \hfill (17.45)

and we have assumed that $\hat{\theta}_n \to^P \theta$, hence

$$(\hat{\theta}_n - \theta) \frac{1}{n} \sum_{i=1}^{n} M(X_i) \to^P 0.$$  \hfill (17.46)

Thus for any $\delta > 0$,

$$P(|\hat{\theta}_n - \theta| < \delta \text{ and } |(\hat{\theta}_n - \theta) \frac{1}{n} \sum_{i=1}^{n} M(X_i)| < \delta] \to 1.$$  \hfill (17.47)

Now take the $\delta < \epsilon$, where $\epsilon$ is from the assumption (17.16). Then

$$|\hat{\theta}_n - \theta| < \delta \implies |\theta^*_n - \theta| < \delta$$
$$\implies |l'''(\theta^*_n; x_i)| \leq M(x_i) \text{ by (17.16)}$$
$$\implies |\frac{1}{n} l'''_n(\theta^*_n)| \leq \frac{1}{n} \sum_{i=1}^{n} M(x_i).$$  \hfill (17.48)

Thus

$$|\hat{\theta}_n - \theta| < \delta \text{ and } |(\hat{\theta}_n - \theta) \frac{1}{n} \sum_{i=1}^{n} M(X_i)| < \delta \implies |(\hat{\theta}_n - \theta) \frac{1}{n} l'''(\theta^*_n)| < \delta,$$  \hfill (17.49)

and (17.47) shows that

$$P[|(\hat{\theta}_n - \theta) \frac{1}{n} l'''(\theta^*_n)| < \delta] \to 1.$$  \hfill (17.50)

That is,

$$(\hat{\theta}_n - \theta) \frac{1}{n} l'''(\theta^*_n) \to^P 0.$$  \hfill (17.51)

Putting together (17.42), (17.44), and (17.51),

$$\sqrt{n} (\hat{\theta}_n - \theta) = -\frac{\sqrt{n} \frac{1}{n} l'_n(\theta)}{\frac{1}{n} l''_n(\theta) + (\hat{\theta}_n - \theta) \frac{1}{n} l'''_n(\theta^*_n) / 2}$$
$$\to^D -\frac{N(0, I_1(\theta))}{-I_1(\theta) + 0}$$
$$= N(0, \frac{1}{I_1(\theta)}),$$  \hfill (17.52)

which proves the theorem (17.40). \hfill $\square$
Chapter 17. Likelihood Estimation

Note. The assumption that we have a consistent sequence of roots can be relaxed to the condition (17.36), that is, \( \hat{\theta}_n \) has to be a root of \( l'_n \) only with high probability:

\[
P[l'(\hat{\theta}_n) = 0 \text{ and } \hat{\theta}_n \in \Theta] \rightarrow 1. \quad (17.53)
\]

As in the exponential family case, if \( I_1(\theta) \) is continuous, \( I_n(\hat{\theta}_n) \rightarrow^P I_1(\theta) \), so that (17.3) holds here, too. It may be that \( I_1(\theta) \) is annoying to calculate. One can instead use the observed Fisher Information, which is defined to be

\[
\hat{I}_1(\theta; x_1, \ldots, x_n) = -\frac{1}{n} l''_n(\theta). \quad (17.54)
\]

The advantage is that the second derivative itself is used, and the expected value of it does not need to be calculated. Using \( \hat{\theta}_n \) yields a consistent estimate of \( I_1(\theta) \):

\[
\hat{I}_1(\hat{\theta}_n; x_1, \ldots, x_n) \rightarrow^P I_1(\theta) + 0, \quad (17.55)
\]

by (17.44) and (17.51). It is thus legitimate to use, for large \( n \), either of the following as approximate 95% confidence intervals:

\[
\hat{\theta}_n \pm 2 \frac{1}{\sqrt{n I_1(\hat{\theta}_n)}} \quad (17.56)
\]

or

\[
\left( \hat{\theta}_n \pm 2 \frac{1}{\sqrt{n \hat{I}_1(\hat{\theta}_n; x_1, \ldots, x_n)}} \right) = \left( \hat{\theta}_n \pm 2 \frac{1}{\sqrt{-l''_n(\hat{\theta}_n)}} \right). \quad (17.57)
\]

Example: Fruit flies. Go back to the fruit fly example in HW#5 of STAT 410. We used Newton-Raphson to calculate the MLE, iterating the equation

\[
\theta^{(i)} = \theta^{(i-1)} - \frac{l'_n(\theta^{(i-1)})}{l''_n(\theta^{(i-1)})}. \quad (17.58)
\]

Starting with the guess \( \theta^{(1)} = 0.5 \), the iterations proceeded as follows:

\[
\begin{array}{c|c|c|c}
   i & \theta^{(i)} & l'_n(\theta^{(i)}) & l''_n(\theta^{(i)}) \\
   \hline
   1 & 0.5 & -5.333333 & -51.55556 \\
   2 & 0.396552 & 0.038564 & -54.50161 \\
   3 & 0.397259 & 0.000024 & -54.43377 \\
   4 & 0.397260 & 0 & -54.43372 \\
\end{array} \quad (17.59)
\]

So \( \hat{\theta}_{\text{MLE}} = 0.3973 \), but notice that we automatically have the observed Fisher Information at the MLE as well,

\[
\hat{I}_1(\hat{\theta}_n; y_1, \ldots, y_{10}) = \frac{1}{10} 54.4338. \quad (17.60)
\]

An approximate 95% confidence interval for \( \theta \) is then

\[
(0.3973 \pm \frac{2}{\sqrt{54.4338}}) = (0.3973 \pm 2 \times 0.1355) = (0.1263, 0.6683). \quad (17.61)
\]

\[
\]
17.4 Asymptotic efficiency

We do not expected the MLE to be unbiased or to have a variance that achieves the Cramér-Rao Lower Bound. In fact, it may be that the mean or variance of the MLE does not exist. For example, the MLE for $1/\lambda$ in the Poisson case is $1/\bar{X}_n$, which does not have a finite mean because there is a positive probability that $\bar{X}_n = 0$. But under the given conditions, if $n$ is large, the MLE is close in distribution to a random variable that is unbiased and that does achieve the CRLB. Recall that we defined efficiency of an unbiased estimator by the ratio of the CRLB to the variance of the estimator. A parallel asymptotic notion, the asymptotic efficiency, looks at the ratio of $1/I_1(\theta)$ to the variance in the asymptotic distribution of appropriate estimators.

A sequence $\delta_n$ is a consistent and asymptotically normal sequence of estimators of $g(\theta)$ if

$$\delta_n \xrightarrow{p} g(\theta) \quad \text{and} \quad \sqrt{n} (\delta_n - g(\theta)) \xrightarrow{D} N(0, \sigma_g^2(\theta)) \quad \tag{17.62}$$

for some $\sigma_g^2(\theta)$. That is, it is consistent and asymptotically normal. The asymptotic normality implies the consistency, because $g(\theta)$ is subtracted from the estimator in the second convergence.

**Theorem 13.** Suppose the conditions in Section 17.1 hold. Then if the sequence $\delta_n$ is a consistent and asymptotically normal estimator of $g(\theta)$, and $g'$ is continuous, then

$$\sigma_g^2(\theta) \geq \frac{g'(\theta)^2}{I_1(\theta)} \quad \tag{17.63}$$

for all $\theta \in \Theta$ except perhaps for a set of Lebesgue measure 0.

That coda about “Lebesgue measure 0” is there because it is possible to trick up the estimator so that it is “superefficient” at a few points. If $\sigma_g^2(\theta)$ is continuous in $\theta$, then you can ignore that bit. Also, the conditions need not be quite as strict as in Section 17.1 in that the part about the third derivative in (17.5) can be dropped, and (17.16) can be changed to be about the second derivative.

**Definition 31.** If the conditions above hold, then the asymptotic efficiency of the sequence $\delta_n$ is

$$\text{AsympEff}_\theta(\delta_n) = \frac{g'(\theta)^2}{I_1(\theta)\sigma_g^2(\theta)}. \quad \tag{17.64}$$

If the asymptotic efficiency is 1, then the sequence is said to be asymptotically efficient.

A couple of immediate consequences follow, presuming the conditions hold.

1. The maximum likelihood estimator is asymptotically efficient, because $\sigma^2(\theta) = 1/I_1(\theta)$.

2. If $\hat{\theta}_n$ is an asymptotically efficient estimator of $\theta$, then $g(\hat{\theta}_n)$ is an asymptotically efficient estimator of $g(\theta)$ by the $\Delta$-method.

**Example.** In the location family case, Mallard problem 3 of HW#3 found the asymptotic efficiencies (which do not depend on $\theta$) of the mean and median in some cases:
Chapter 17. Likelihood Estimation

<table>
<thead>
<tr>
<th>$f$</th>
<th>AsympEff(Mean)</th>
<th>AsympEff(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>1</td>
<td>0.6366</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.9119</td>
<td>0.75</td>
</tr>
<tr>
<td>Cauchy</td>
<td>0</td>
<td>0.8106</td>
</tr>
<tr>
<td>Double Exponential</td>
<td>0.5</td>
<td>1</td>
</tr>
</tbody>
</table>

The assumptions actually do not hold for the Double Exponential density, but the asymptotic efficiency calculations are still valid. For these cases, the MLE is asymptotically efficient; in the Normal case the MLE is the mean, and in the Double Exponential case the MLE is the median. Also, although we do not prove it, the Pitman Estimator is asymptotically efficient, which should make sense because its variance is the lowest possible. It is a matter of showing that it is asymptotically normal.

Lehmann (1991) in Table 4.4 has more calculations for the asymptotic efficiencies of some trimmed means. Here are some of the values, where $\alpha$ is the trimming proportion (the mean is $\alpha = 0$ and the median is $\alpha = 1/2$):

<table>
<thead>
<tr>
<th>$f \downarrow$ ; $\alpha \rightarrow$</th>
<th>0</th>
<th>1/8</th>
<th>1/4</th>
<th>3/8</th>
<th>1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>1.00</td>
<td>0.94</td>
<td>0.84</td>
<td>0.74</td>
<td>0.64</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.91</td>
<td>0.99</td>
<td>0.95</td>
<td>0.86</td>
<td>0.75</td>
</tr>
<tr>
<td>t$_5$</td>
<td>0.80</td>
<td>0.99</td>
<td>0.96</td>
<td>0.88</td>
<td>0.77</td>
</tr>
<tr>
<td>t$_3$</td>
<td>0.50</td>
<td>0.96</td>
<td>0.98</td>
<td>0.92</td>
<td>0.81</td>
</tr>
<tr>
<td>Cauchy</td>
<td>0.00</td>
<td>0.50</td>
<td>0.79</td>
<td>0.88</td>
<td>0.81</td>
</tr>
<tr>
<td>Double Exponential</td>
<td>0.50</td>
<td>0.70</td>
<td>0.82</td>
<td>0.91</td>
<td>1.00</td>
</tr>
</tbody>
</table>

This table can help you choose what trimming amount you would want to use, depending on what you think your $f$ might be. You can see that between the mean and a small amount of trimming (1/8), the efficiencies of most distributions go up substantially, while the Normal’s goes down only a small amount. With 25% trimming, all have at least a 79% efficiency.

17.5 Multivariate parameters

The work so far assumed that $\theta$ was one-dimensional (although the data could be multidimensional). Everything follows for multidimensional parameters $\bar{\theta}$, with some extended definitions. Now assume that $\Theta \subset \mathbb{R}^K$, and that $\Theta$ is open. The derivative of the loglikelihood is now $K$-dimensional,

$$\nabla l_n(\theta) = \sum_{i=1}^{n} \nabla l(\theta; x_i), \quad (17.67)$$

where

$$\nabla l(\theta; x) = \left( \frac{\partial \log(f(x | \theta))}{\partial \theta_1} \right) \ldots \left( \frac{\partial \log(f(x | \theta))}{\partial \theta_K} \right). \quad (17.68)$$

---

3 Theory of Point Estimation, Wadsworth. This is the first edition of the Lehmann-Casella book.
17.5. Multivariate parameters

The MLE then satisfies the equations

$$\nabla l_n(\hat{\theta}_n) = 0.$$ (17.69)

The score function is also a $K \times 1$ vector, in fact,

$$S(\theta; x) = \nabla l(\theta; x),$$ (17.70)

and the Fisher information in one observation is a $K \times K$ matrix,

$$I_1(\theta) = \text{Cov}_\theta[S(\theta)]$$ (17.71)

$$= -E \begin{bmatrix} \frac{\partial^2 \log(f(x | \theta))}{\partial \theta_1^2} & \frac{\partial^2 \log(f(x | \theta))}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \log(f(x | \theta))}{\partial \theta_1 \partial \theta_K} \\ \frac{\partial^2 \log(f(x | \theta))}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log(f(x | \theta))}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \log(f(x | \theta))}{\partial \theta_2 \partial \theta_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log(f(x | \theta))}{\partial \theta_K \partial \theta_1} & \frac{\partial^2 \log(f(x | \theta))}{\partial \theta_K \partial \theta_2} & \cdots & \frac{\partial^2 \log(f(x | \theta))}{\partial \theta_K^2} \end{bmatrix}. $$ (17.72)

I won’t detail all the assumptions, but they are basically the same as before, except that they apply to all the partial and mixed partial derivatives. The equation (17.11),

$$0 < I_1(\theta) < \infty \quad \text{for all } \theta \in \Theta,$$ (17.73)

means that $I_1(\theta)$ is positive definite, and all its elements are finite. The two main results are next:

1. If $\hat{\theta}_n$ is a consistent sequence of roots of the derivative of the loglikelihood, then

$$\sqrt{n} (\hat{\theta}_n - \theta) \rightarrow^D N(0, I^{-1}(\theta)).$$ (17.74)

2. If $\delta_n$ is a consistent and asymptotically normal sequence of estimators of $g(\theta)$, where the partial derivatives of $g$ are continuous, and

$$\sqrt{n} (\delta_n - g(\theta)) \rightarrow^D N(0, \sigma^2_g(\theta)),$$ (17.75)

then

$$\sigma^2_g(\delta) \geq D_g(\theta)'I_1^{-1}(\theta)D_g(\theta)$$ (17.76)

for all $\theta \in \Theta$ (except possibly for a few), where $D_g$ is the $K \times 1$ vector of partial derivatives $\partial g(\theta)/\partial \theta_i$ as in the multivariate $\Delta$-method in (16.116).

The lower bound in (17.76) is the variance in the asymptotic distribution of $\sqrt{n}(g(\hat{\theta}_n) - g(\theta))$, where $\hat{\theta}_n$ is the MLE of $\theta$. Which is to say that the MLE of $g(\theta)$ is asymptotically efficient.

**Example: Normal**

Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$. The MLE of $\theta = (\mu, \sigma^2)'$ is $\hat{\theta} = (\bar{X}_n, S_n^2)'$, where the $S_n^2$ is divided by $n$. To find the information matrix, start with

$$l(\mu, \sigma^2; x) = -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} - \frac{1}{2} \log(\sigma^2).$$ (17.77)
Then
\[ S(\mu, \sigma^2; x) = \left( \frac{x-\mu}{\sigma^2} \right) \cdot \left( \frac{1}{2} \frac{(x-\mu)^2}{\sigma^4} - \frac{1}{2\sigma^2} \right). \] (17.78)

(For fun you can check that \( E[S] = 0 \).) To illustrate, we’ll use two methods for finding the information matrix. First,
\[
I_1(\theta) = \text{Cov}_\theta[S(\mu, \sigma^2; X)]
= \begin{pmatrix}
E[(X-\mu)^2] & E\left[\frac{1}{2} \frac{(X-\mu)^3}{\sigma^6}\right] \\
E\left[\frac{1}{2} \frac{(X-\mu)^2}{\sigma^4}\right] & \text{Var}\left[\frac{1}{2} \frac{(X-\mu)^2}{\sigma^4}\right]
\end{pmatrix} (Z = \frac{X-\mu}{\sigma} \sim N(0,1))
= \begin{pmatrix}
\frac{1}{\sigma^2} & 0 \\
0 & \frac{1}{2} \frac{1}{\sigma^4}
\end{pmatrix}.
\]

Using the second derivatives,
\[
I_1(\theta) = -E \left[ \begin{pmatrix}
-\frac{1}{\sigma^2} & \frac{X-\mu}{\sigma^4} \\
\frac{X-\mu}{\sigma^4} & \frac{(X-\mu)^3}{\sigma^6} + \frac{1}{2\sigma^2}
\end{pmatrix} \right]
= -E \left[ \begin{pmatrix}
-\frac{1}{\sigma^2} & 0 \\
0 & \frac{1}{2} \frac{1}{\sigma^4}
\end{pmatrix} \right]
= \begin{pmatrix}
\frac{1}{\sigma^2} & 0 \\
0 & \frac{1}{2} \frac{1}{\sigma^4}
\end{pmatrix}.
\]

Good, they both gave the same answer! Fortunately, the information is easy to invert, hence
\[
\sqrt{n} \left( \begin{pmatrix}
\bar{X}_n \\
S_n^2
\end{pmatrix} - \left( \begin{pmatrix}
\mu \\
\sigma^2
\end{pmatrix}
\right) \right) \rightarrow^D N(0_2, \sigma^2 \begin{pmatrix}
0 & 0 \\
0 & 2\sigma^4
\end{pmatrix}). \] (17.79)

which is what we found in (16.126) using the \( \Delta \)-method.
Chapter 18

Hypothesis Testing

Estimation addresses the question, “What is $\theta$?” Hypothesis testing addresses questions like, “Is $\theta = 0$?” Confidence intervals do both. It will give a range of plausible values, and if you wonder whether $\theta = 0$ is plausible, you just check whether 0 is in the interval. But hypothesis testing also address broader questions in which confidence intervals may be clumsy. Some types of questions for hypothesis testing:

- Is a particular drug more effective than a placebo?
- Are cancer and smoking related?
- Is the relationship between amount of fertilizer and yield linear?
- Is the distribution of income the same among men and women?
- In a regression setting, are the errors independent? Normal? Homoskedastic?
- Does God exist?  

The main feature of hypothesis testing problems is that there are two competing models under consideration, the null hypothesis model and the alternative hypothesis model. The random variable (vector) $X$ and space $X$ are the same in both models, but the sets of distributions are different, being denoted $P_0$ and $P_A$ for the null and alternative, respectively, where $P_0 \cap P_A = \emptyset$. If $P$ is the probability distribution for $X$, then the hypotheses are written

$$H_0 : P \in P_0 \text{ versus } H_A : P \in P_A.$$  \hspace{1cm} (18.1)

Often both models will be parametric:

$$P_0 = \{P_\theta \mid \theta \in \Theta_0 \} \text{ and } P_A = \{P_\theta \mid \theta \in \Theta_A \}, \text{ with } \Theta_0, \Theta_A \subset \Theta, \Theta_0 \cap \Theta_A = \emptyset,$$ \hspace{1cm} (18.2)

for some overall parameter space $\Theta$. It is not unusual, but also not required, that $\Theta_A = \Theta - \Theta_0$. In a parametric setting, the hypotheses are written

$$H_0 : \theta \in \Theta_0 \text{ versus } H_A : \theta \in \Theta_A.$$ \hspace{1cm} (18.3)

---

1The chance is 67%, according to http://education.guardian.co.uk/higher/sciences/story/0,12243,1164894,00.html
Some examples:

1. Suppose $n_D$ people are given a drug, and $n_P$ are given a placebo, and in each group the number of people who improve is counted. The data could be modeled as $X_D \sim \text{Binomial}(n_D, p_D)$ and $X_P \sim \text{Binomial}(n_P, p_P)$, $X_D$ and $X_P$ independent. The question is whether the drug worked, hence we would test

$$H_0 : p_D = p_P \text{ versus } H_A : p_D > p_P. \quad (18.4)$$

More formally, these would be written as in (18.3) as

$$H_0 : (p_D, p_P) \in \{(u_D, u_P) \mid 0 < u_D = u_P < 1\} \text{ versus } \quad (18.5)$$
$$H_A : (p_D, p_P) \in \{(u_D, u_P) \mid 0 < u_P < u_D < 1\}. \quad (18.6)$$

The overall parameter space is $(0,1) \times (0,1)$.

2. A simple linear regression model has $Y_i = \alpha + \beta x_i + E_i$, $i = 1, \ldots, n$, where the $x_i$’s are fixed and known, and the $E_i$’s are iid. Often, it is assumed the errors are also $N(0, \sigma_e^2)$. Is that assumption plausible? One can test

$$H_0 : E_1, \ldots, E_n \text{ iid } N(0, \sigma_e^2), \sigma_e^2 > 0 \text{ versus } H_A : E_1, \ldots, E_n \text{ iid } F, \ F \neq N(0, \sigma_e^2).$$

To write this more formally, we have the parameters $(\alpha, \beta, F)$, where $F$ is the common distribution function of the $E_i$’s. Then

$$\Theta_0 = \mathbb{R}^2 \times \{N(0, \sigma_e^2) \mid \sigma_e^2 > 0\}, \quad (18.8)$$

and

$$\Theta_A = \mathbb{R}^2 \times \{\text{all distribution functions } F \} - \Theta_0, \quad (18.9)$$

so that we test

$$H_0 : (\alpha, \beta, F) \in \Theta_0 \text{ versus } H_A : (\alpha, \beta, F) \in \Theta_A. \quad (18.10)$$

The alternative model is semi-parametric in that part of the parameter, $\alpha$ and $\beta$, lies in a finite-dimensional space, and part, the $F$, lies in an infinite-dimensional space.

3. Another goodness-of-fit testing problem for regression asks whether the mean is linear. That is, the model is $Y_i = g(x_i) + E_i$, $E_1, \ldots, E_n$ iid $N(0, \sigma_e^2)$, and the null hypothesis is that $g(x)$ is a linear function:

$$H_0 : g(x) = \alpha + \beta x_i, (\alpha, \beta) \in \mathbb{R}^2 \text{ versus } \quad (18.11)$$

$$H_A : g \in \{h : \mathbb{R} \to \mathbb{R} \mid h''(x) \text{ is continuous } \} - \{\alpha + \beta x \mid (\alpha, \beta) \in \mathbb{R}^2\}.$$  

More formally, we have that

$$Y_i \sim N(g(x_i), \sigma_e^2), i = 1, \ldots, n,$$

$$\Theta_0 = \{h \mid h(x) = \alpha + \beta x, (\alpha, \beta) \in \mathbb{R}^2\} \times (0, \infty),$$

$$\Theta_A = \{(h, \sigma_e^2) \mid h : \mathbb{R} \to \mathbb{R}, h''(x) \text{ is continuous } \} \times (0, \infty) - \Theta_0,$$

and

$$H_0 : (g, \sigma_e^2) \in \Theta_0 \text{ versus } H_A : (g, \sigma_e^2) \in \Theta_A. \quad (18.12)$$

The alternative is again semi-parametric.
Mathematically, there is no particular reason to designate one of the hypotheses null and the other alternative. In practice, the null hypothesis tends to be the one that represents the status quo, or that nothing unusual is happening, or that everything is ok, or that the new isn’t any better than the old, or that the defendant is innocent. The null also tends to be simpler (lower-dimensional), although that is not necessarily the case, e.g., one could test \( H_0 : p_D \leq p_P \) versus \( H_A : p_D > p_P \) in (18.4).

There are several approaches to testing the hypotheses, e.g., ...

1. **Accept/Reject or fixed \( \alpha \) or Neyman-Pearson.** This approach is action-oriented: Based on the data \( x \), you either accept or reject the null hypothesis. One drawback is that whatever decision you make, there is not an obvious way to assess the accuracy of your decision. (In estimation, you would hope to have an estimate and its standard error.) Often the size or level is cited as the measure of accuracy, which is the chance of a false positive, i.e., rejecting the null when the null is true. It is not a very good measure. It really gives the long-run chance of making an error if one kept doing hypothesis testing with the same \( \alpha \). That is, it refers to the infinite tests you haven’t done yet, not the one you are analyzing.

2. **Evidence against the null.** Rather than make a decision, this approach produces some number measuring support, or lack of it, for the null based on \( x \). Possible such measures include

   a) **p-value.** Given the data \( x \), the p-value is the chance, if the null hypothesis were true, that one would see \( X \) as extreme as, or more extreme than, \( x \). The smaller, the more evidence against the null. The definition is slightly convoluted, and many believe that the p-value is the chance that the null hypothesis is true. Not so!

   b) **Bayes factor or likelihood ratio.** When both hypotheses are simple, i.e., single points, say \( \theta_0 \) and \( \theta_A \), respectively, then, given \( x \), the ratio of the likelihoods is

\[
LR(x) = \frac{f_{\theta_A}(x)}{f_{\theta_0}(x)}. \tag{18.13}
\]

The larger the LR is, the more likely the alternative is relative to the null, hence measures evidence against the null. When the hypotheses are not simple, the likelihoods are averaged before taking the ratio.

   c) **Probability the null is true.** One needs a prior distribution on \( \Theta_0 \cup \Theta_A \), but given that, say \( \pi \), it is straightforward to find the probability of the null:

\[
P[H_0 \mid X = x] = \frac{\int_{\Theta_0} f(x \mid \theta) \pi(\theta) d\theta}{\int_{\Theta_0 \cup \Theta_A} f(x \mid \theta) \pi(\theta) d\theta}. \tag{18.14}
\]

This measure of evidence is the most direct and easily interpreted, but is quite sensitive to the prior distribution.

None of these approaches is entirely satisfactory, unless you really do have a prior you believe in. Notice that any of the evidentiary methods can be turned into an action, e.g., rejecting the null if the p-value is less than something, or the likelihood ratio is larger than something, or the probability of the null is less than something. Now to the details of implementation.
18.1 Accept/Reject

There is a great deal of terminology associated with the accept/reject paradigm, but the basics are fairly simple. Start with a test statistic \( T(x) \), which is a function \( T: \mathcal{X} \to \mathbb{R} \) that measures in some sense the difference between the data \( x \) and the null hypothesis. For example, suppose the data are \( X_1, \ldots, X_n \) iid \( N(\mu, \sigma^2) \), and we wish to test

\[
H_0 : \mu = \mu_0, \sigma^2 > 0 \quad \text{versus} \quad H_A : \mu \neq \mu_0, \sigma^2 > 0 \tag{18.15}
\]

for some fixed known \( \mu_0 \). The “\( \sigma^2 > 0 \)” in both hypotheses is the technical way of saying, “\( \sigma^2 \) is unknown.” The test statistic should measure the closeness of \( \bar{x} \) to \( \mu_0 \). For example, the usual two-sided \( t \)-statistic,

\[
T(x) = \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}}. \tag{18.16}
\]

The absolute value is there because sample means far above and far below \( \mu_0 \) both favor the alternative. This \( T \) is not the only plausible test statistic. The sign test takes the point of view that if the null hypothesis were true, then about half the observation would be above \( \mu_0 \), and about half below. Thus if the proportion below \( \mu_0 \) is too far from \( \frac{1}{2} \), we would worry about the null. The two-sided sign test statistic is

\[
S(x) = \left\lfloor \frac{\#\{x_i < \mu_0\}}{n} \right\rfloor - \frac{1}{2}. \tag{18.17}
\]

Next, choose a cutoff point \( c \) which represents how large the test statistic can be before rejecting the null hypothesis. That is,

\[
\text{the test} \left\{ \begin{array}{l}
\text{Rejects the null} \quad \text{if} \quad T(x) > c \\
\text{Accepts the null} \quad \text{if} \quad T(x) \leq c
\end{array} \right.. \tag{18.18}
\]

The bigger the \( c \), the harder it is to reject the null.

In practice, one usually doesn’t want to reject the null unless there is substantial evidence against it. The situation is similar to the courts in a criminal trial. The defendant is “presumed innocent until proven guilty.” That is, the jury imagines the defendant is innocent, then considers the evidence, and only if the evidence piles up so that the jury believes the defendant is “guilty beyond reasonable doubt” does it convict. The accept/reject approach to hypothesis testing parallels the courts with the following connections:

<table>
<thead>
<tr>
<th>Courts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defendant innocent</td>
</tr>
<tr>
<td>Evidence</td>
</tr>
<tr>
<td>Defendant declared guilty</td>
</tr>
<tr>
<td>Defendant declared not guilty</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null hypothesis true</td>
</tr>
<tr>
<td>Data</td>
</tr>
<tr>
<td>Reject the null</td>
</tr>
<tr>
<td>Accept the null</td>
</tr>
</tbody>
</table>

(18.19)

Notice that the jury does not say that the defendant is innocent, but either guilty or not guilty. “Not guilty” is a way to say that there is not enough evidence to convict, not to say the jury is confident the defendant is innocent. Similarly, in hypothesis testing accepting the hypothesis does not mean one is confident it is true. Rather, it may be true, or just that there is not enough evidence to reject it.

“Reasonable doubt” is quantified by looking at the level of the test:
**Definition 32.** A hypothesis test (18.18) has level $\alpha$ if

$$P_{\theta}[T(X) > c] \leq \alpha \text{ for all } \theta \in \Theta_0.$$  

(18.20)

That is, the chance of rejecting the null when it is true is no larger than $\alpha$. (So the chance of convicting an innocent person is less than $\alpha$.) Note that a test with level 0.05 also has level 0.10. A related concept is the size of a test, which is the smallest $\alpha$ for which it is level $\alpha$.

It is good to have a small $\alpha$, because that means one is unlikely to reject the null erroneously. Of course, you could take $c = \infty$, and never reject, so that $\alpha = 0$, and you would never reject the null erroneously. But if the null were false, you’d never reject, either, which is bad. That is, there are two types of error to balance. These errors are called, rather colorlessly, Type I and Type II errors:

<table>
<thead>
<tr>
<th>Truth</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>Accept $H_0$</td>
</tr>
<tr>
<td>$H_A$</td>
<td>Type I error</td>
</tr>
</tbody>
</table>

(18.21)

It would have been better if the terminology had been in line with medical and other usage, where a *false positive* is rejecting when the null is true, e.g., saying you have cancer when you don’t, and *false negative*, e.g., saying everything is ok when it is not.

Anyway the probability of Type I error is bounded above by the level. Conventionally, one chooses the level $\alpha$, then finds the cutoff point $c$ to yield that level. Then the probability of Type II is whatever it is. Traditionally, more emphasis is on the power than the probability of Type II error, where $\text{Power}_{\theta} = 1 - P_{\theta}[\text{Type II Error}]$ when $\theta \in \Theta_A$. Power is good.

**Example.** Consider the sign statistic in (18.17), where the cutoff point is $c = 1/5$, so the test rejects the null hypothesis if

$$S(x) = \left| \frac{\# \{X_i < \mu_0\}}{n} - \frac{1}{2} \right| > \frac{1}{5}. \quad (18.22)$$

To find the level and power, we need the distribution of $S$. Now

$$Y \equiv \# \{X_i < \mu_0\} \sim \text{Binomial}(n, P[X_i < \mu_0])$$  

(18.23)

and

$$P[X_i < \mu_0] = P[N(0, 1) < (\mu_0 - \mu)/\sigma] = \Phi[(\mu_0 - \mu)/\sigma], \quad (18.24)$$

hence

$$Y \sim \text{Binomial}(n, \Phi[(\mu_0 - \mu)/\sigma]).$$  

(18.25)

Suppose $n = 25$. For the level, we assume the null hypothesis in (18.15) is true, so that $\mu = \mu_0$. Then $Y \sim \text{Binomial}(25, \frac{1}{2})$. Note that the distribution of $Y$, hence $S$, is the same for all $\sigma^2$'s, and long as $\mu = \mu_0$. Thus

$$\alpha = P\left[ \left| \frac{Y}{25 - \frac{1}{2}} \right| > \frac{1}{5} \right] = P[|Y - 12.5| > 5] = P[Y \leq 7 \text{ or } Y \geq 18] = 2(0.02164) = 0.0433.$$  

(18.26)
The power depends on \( \mu, \mu_0 \) and \( \sigma \); actually, on the scaled difference between \( \mu \) and \( \mu_0 \):

\[
Y \sim \text{Binomial}(n, \Phi(\Delta)), \quad \Delta = \frac{\mu_0 - \mu}{\sigma}.
\] (18.27)

The \( \Delta \) is a measure of how far away from the null the \((\mu, \sigma^2)\) is. Some powers are calculated below, as well as the level again.

Note that the power is symmetric in that \( \Delta \) and \(-\Delta \) have the same power. Here is a plot, often called a power curve, although when \( \Delta = 0 \), it plots the level \( \alpha \):

![Power of Sign Test](image)

Lowering the cutoff point \( c \) will increase the power (good) and level (bad), while increasing \( c \) will decrease power (bad) but also the level (good). So there is a tradeoff.

The next goal is to find a “good” test, that is, one with low level and high power. The Neyman-Pearson approach is to specify \( \alpha \) and try to find the level \( \alpha \) test with the best power. There may not be one best test for all \( \theta \in \Theta_A \), which makes the problem harder. It is analogous to trying to find the unbiased estimator with the lowest variance. The next section introduces test functions and randomized tests, both of which are mainly to make the theory flow better. In practice, tests as in (18.18) are used.
18.2 Test functions

Suppose in the sign test example (18.22), we want $\alpha = 0.05$. The test we used, with $c = 1/5$, does have level $\alpha$, since the probability of rejecting under the null, the size, is 0.0433, which is less than $\alpha$. But we can obtain better power if we increase the size to 0.05. Thus we want to find $c$ so that, from (18.26),

$$P[|Y - 12.5| > 25c] = 0.05, \quad Y \sim \text{Binomial}(25, \frac{1}{2}). \quad (18.29)$$

Let’s decrease $c$ a bit. If $c = 0.9/5$, then

$$P[|Y - 12.5| > 25 \times 0.9/5] = P[Y \leq 7 \text{ or } Y \geq 18] = 2(0.02164) = 0.0433, \quad (18.30)$$

the same as for $c = 1/5$, but if we make $c$ even a tiny bit smaller than 0.9/5, say $c = 0.9/5 - \epsilon$ for tiny $\epsilon > 0$, then

$$P[|Y - 12.5| > 25 \times (0.9/5 - \epsilon)] = P[Y \leq 8 \text{ or } Y \geq 17] = 2(0.05388) = 0.1078. \quad (18.31)$$

But that is too large, which means we cannot choose a $c$ to give exactly the $\alpha$ we want. Unless we use a randomized test, which means that if the test statistic equals the cutoff point, then we reject the null with a probability that may be strictly between 0 and 1. Thus the test (18.18) is generalized slightly to

$$\text{the test } \left\{ \begin{array}{ll}
\text{Rejects the null} & \text{if } T(x) > c \\
\text{Rejects the null with probability } \gamma(x) & \text{if } T(x) = c \\
\text{Accepts the null} & \text{if } T(x) < c
\end{array} \right., \quad (18.32)$$

where $\gamma(x)$ is a function with $0 \leq \gamma(x) \leq 1$.

For the sign test, we’ll take $c = 0.9/5$, and let $\gamma$ be a constant, so that

$$\text{the test } \left\{ \begin{array}{ll}
\text{Rejects the null} & \text{if } \left| \frac{Y}{n} - \frac{1}{2} \right| > \frac{0.9}{5} \\
\text{Rejects the null with probability } \gamma & \text{if } \left| \frac{Y}{n} - \frac{1}{2} \right| = \frac{0.9}{5} \\
\text{Accepts the null} & \text{if } \left| \frac{Y}{n} - \frac{1}{2} \right| < \frac{0.9}{5}
\end{array} \right. \quad (18.33)$$

Then

$$P_{\mathbb{E}}[\text{The test rejects } H_0] = \frac{\gamma}{n} \mathbb{I}_{\left| \frac{Y}{n} - \frac{1}{2} \right| > \frac{0.9}{5}} + \gamma \frac{\mathbb{I}_{\left| \frac{Y}{n} - \frac{1}{2} \right| = \frac{0.9}{5}}}{n}$$

$$= P_{\mathbb{E}}[Y \leq 7 \text{ or } Y \geq 18] + \gamma P_{\mathbb{E}}[Y = 8 \text{ or } Y = 17]. \quad (18.34)$$

To find $\gamma$, we do the calculations under the null, so that $Y \sim \text{Binomial}(25, \frac{1}{2})$:

$$P_{H_0}[\text{The test rejects } H_0] = 0.0433 + \gamma 0.0645. \quad (18.35)$$

Taking $\gamma = 0$ yields the size 0.0433 test, and taking $\gamma = 1$, we have the size 0.1078 test from (18.31). To get 0.05, solve

$$0.05 = 0.0433 + \gamma 0.0645 \implies \gamma = 0.1042. \quad (18.36)$$
Which means that if the sign statistic is greater than \( \frac{0.9}{5} \), you reject, and if the sign statistic is equal to \( \frac{0.9}{5} \), you flip a biased coin with probability of heads being 0.1039, and reject if it comes up heads.

\[
\begin{array}{c|c|c}
\Delta & P[\text{Reject } H_0] & P[\text{Reject } H_0] \\
\gamma = 0 & \gamma = 0.1042 \\
\hline
-1.5 & 0.9998 & 0.9999 \\
-1.0 & 0.9655 & 0.9679 \\
-0.5 & 0.4745 & 0.4920 \\
0 & 0.0433 & 0.0500 \\
0.5 & 0.4745 & 0.4920 \\
1.0 & 0.9655 & 0.9679 \\
1.5 & 0.9998 & 0.9999 \\
\end{array}
\]

(18.37)

So this randomized test is also level \( \alpha \), but has slightly better power.

Randomized tests are nice theoretically, but make sure you don’t talk about them outside of this course. Especially to applied statisticians. They’ll blow their top. Which makes sense, because after they do all the work of conducting an experiment to test whether their drug works, they don’t want the FDA to flip a coin to make the decision.

A convenient way to express a (possibly) randomized test is to use “1” to mean reject and “0” to mean accept, and keep the \( \gamma \) for the randomized part. That is, for test (18.32), define the function \( \phi \) by

\[
\phi(x) = \begin{cases} 
1 & \text{if } T(x) > c \\
\gamma(x) & \text{if } T(x) = c \\
0 & \text{if } T(x) < c 
\end{cases}
\]

(18.38)

Now the power and level are just expected values of the \( \phi \) because

\[
P_\theta[\text{The test rejects } H_0] = E_\theta[\phi(X)].
\]

(18.39)

**Definition 33.** A test function is a function \( \phi : X \rightarrow [0, 1] \), with the interpretation that the test rejects \( H_0 \) with probability \( \phi(x) \) if \( x \) is observed.

Now to find good tests.

### 18.3 Simple versus simple

We will start simple, where each hypothesis has exactly one distribution, so that we test

\[
H_0 : \theta = \theta_0 \text{ versus } H_A : \theta = \theta_A,
\]

(18.40)

where \( \theta_0 \) and \( \theta_A \) are two different parameter values.

For example, suppose \( X \sim Binomial(4, p) \), and the hypotheses are

\[
H_0 : p = \frac{1}{2} \text{ versus } H_A : p = \frac{3}{5}.
\]

(18.41)

Let \( \alpha = 0.35 \), so that the objective is to find a test \( \phi \) that

\[
\text{Maximizes } E_{3/5}[\phi(X)] \text{ subject to } E_{1/2}[\phi(X)] \leq \alpha = 0.35,
\]

(18.42)
that is, maximizes the power subject to begin of level \( \alpha \). What should we look for? The table below gives the likelihoods under the null and alternative:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( f_{1/2}(x) )</th>
<th>( f_{3/5}(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0625</td>
<td>0.0256</td>
</tr>
<tr>
<td>1</td>
<td>0.2500</td>
<td>0.1536</td>
</tr>
<tr>
<td>2</td>
<td>0.3750</td>
<td>0.3456</td>
</tr>
<tr>
<td>3</td>
<td>0.2500</td>
<td>0.3456</td>
</tr>
<tr>
<td>4</td>
<td>0.0625</td>
<td>0.1296</td>
</tr>
</tbody>
</table>

One could focus on the null hypothesis. If the null is true, then it is more likely that \( X = 2 \), or 1 or 3, than 0 or 4. Consider rejecting the null when \( f_{1/2}(x) \) is small. We would reject when \( X = 0 \) or 4, but that gives probability of rejection of only 0.125. To get up to 0.35, we would need to reject at least part of the time if \( X = 1 \) or 3. The test is then

\[
\phi_1(x) = \begin{cases} 
1 & \text{if } x = 0, 4 \\
0.45 & \text{if } x = 1, 3 \\
0 & \text{if } x = 2 
\end{cases}
\]  

(18.44)

The \( \gamma = 0.45 \) gives the level \( \alpha \). Thus

\[
E_{1/2}[\phi_1(X)] = 0.125 + 0.45 \times 0.5 = 0.35 \quad \text{and}
\]

\[
Power = E_{3/5}[\phi_1(X)] = 0.0256 + 0.1296 + 0.45 \times (0.1536 + 0.3456) = 0.3798.
\]  

(18.45)

Alternatively, one could look at the alternative density, and reject the null if it is likely that the alternative is true. The values \( X = 2 \) and 3 are most likely, and add to more than 0.35, so one could contemplate the test

\[
\phi_2(x) = \begin{cases} 
0.56 & \text{if } x = 2, 3 \\
0 & \text{if } x = 0, 1, 4 
\end{cases}
\]  

(18.46)

Then

\[
E_{1/2}[\phi_2(X)] = 0.35 \quad \text{and}
\]

\[
Power = E_{3/5}[\phi_2(X)] = 0.3871.
\]  

(18.47)

This second test has a slightly better power than the first, 0.3871 to 0.3798, but neither are much higher than the level \( \alpha = 0.35 \). Is that the best we can do?

**Bang for the buck.** Imagine you have some bookshelves you wish to fill up as cheaply as possible, e.g., to use as props in a play. You do not care about the quality of the books, just their widths (in inches) and prices (in dollars). You have $3.50, and five books to choose from:

<table>
<thead>
<tr>
<th>Book</th>
<th>Cost</th>
<th>Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.625</td>
<td>0.256</td>
</tr>
<tr>
<td>1</td>
<td>2.50</td>
<td>1.536</td>
</tr>
<tr>
<td>2</td>
<td>3.75</td>
<td>3.456</td>
</tr>
<tr>
<td>3</td>
<td>2.50</td>
<td>3.456</td>
</tr>
<tr>
<td>4</td>
<td>0.625</td>
<td>1.296</td>
</tr>
</tbody>
</table>

(18.48)

You are allowed to split the books lengthwise, and pay proportionately. You want to maximize the total number of inches for your $3.50. Then, for example, book 4 is a
better deal than book 0, because they cost the same but book 4 is wider. Also, book 3 is more attractive than book 2, because they are the same width but book 3 is cheaper. Which is better between books 3 and 4? Book 4 is cheaper by the inch: It costs about 48¢ per inch, while book 3 is about 73¢ per inch. This suggests the strategy should be to buy the books that give you the most inches per dollar:

<table>
<thead>
<tr>
<th>Book</th>
<th>Cost</th>
<th>Width</th>
<th>Inches/Dollar</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.625</td>
<td>0.256</td>
<td>0.4096</td>
</tr>
<tr>
<td>1</td>
<td>2.50</td>
<td>1.536</td>
<td>0.6144</td>
</tr>
<tr>
<td>2</td>
<td>3.75</td>
<td>3.456</td>
<td>0.9216</td>
</tr>
<tr>
<td>3</td>
<td>2.50</td>
<td>3.456</td>
<td>1.3824</td>
</tr>
<tr>
<td>4</td>
<td>0.625</td>
<td>1.296</td>
<td>2.0736</td>
</tr>
</tbody>
</table>

(18.49)

Let us definitely buy book 4, and book 3. That costs us $3.125, and gives us 1.296+3.456 = 4.752 inches. We still have 37.5¢ left, with which we can buy a tenth of book 2, giving us another 0.3456 inches, totaling 5.0976 inches.

Returning to the hypothesis testing problem, we can think of having $\alpha$ to spend, and we wish to spend where we get the most “bang for the buck.” Here, bang is power. The key is to look at the likelihood ratio of the densities:

$$LR(x) = \frac{f_{3/5}(x)}{f_{1/2}(x)},$$

(18.50)

i.e.,

<table>
<thead>
<tr>
<th>x</th>
<th>$f_{1/2}(x)$</th>
<th>$f_{3/5}(x)$</th>
<th>LR(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0625</td>
<td>0.0256</td>
<td>0.4096</td>
</tr>
<tr>
<td>1</td>
<td>0.2500</td>
<td>0.1536</td>
<td>0.6144</td>
</tr>
<tr>
<td>2</td>
<td>0.3750</td>
<td>0.3456</td>
<td>0.9216</td>
</tr>
<tr>
<td>3</td>
<td>0.2500</td>
<td>0.3456</td>
<td>1.3824</td>
</tr>
<tr>
<td>4</td>
<td>0.0625</td>
<td>0.1296</td>
<td>2.0736</td>
</tr>
</tbody>
</table>

(18.51)

If $LR(x)$ is large, then the alternative is much more likely than the null is. If $LR(x)$ is small, the null is more likely. One uses the likelihood ratio as the statistic, and finds the right cutoff point. The likelihood ratio test is then

$$\phi_{LR}(x) = \begin{cases} 
  1 & \text{if } LR(x) > c \\
  \gamma & \text{if } LR(x) = c \\
  0 & \text{if } LR(x) < c 
\end{cases}.$$ 

(18.52)

Looking at the table (18.51), we see that taking $c = LR(2)$ works, because we reject if $x = 3$ or 4, and use up only .3125 of our $\alpha$. Then the rest we put on $x = 2$, the $\gamma = 1/10$ since we have .35 – .3125 = 0.0375 left. Thus the test is

$$\phi_{LR}(x) = \begin{cases} 
  1 & \text{if } LR(x) > 0.9216 \\
  0.1 & \text{if } LR(x) = 0.9216 \\
  0 & \text{if } LR(x) < 0.9216 
\end{cases}.$$ 

(18.53)

The last expression is easier to deal with, and valid since $LR(x)$ is a strictly increasing function of $x$. Then the power and level are

$$E_{1/2}[\phi_{LR}(X)] = 0.35 \quad \text{and}$$

$$\text{Power} = E_{3/5}[\phi_{LR}(X)] = 0.1296 + 0.3456 + 0.1 \times 0.3456 = 0.50976.$$ 

(18.54)
This is the same as for the books: The power is identified with the number of inches. Also, the likelihood ratio test is much better than \( \phi_1 \) (18.44) and \( \phi_2 \) (18.46), with powers around 0.38. Is it the best? Yes, that is the Neyman-Pearson Lemma.

### 18.4 The Neyman-Pearson Lemma

Let \( X \) be the random variable or vector with density \( f \), and \( f_0 \) and \( f_A \) be two possible densities for \( X \). We are interested in testing

\[
H_0 : f = f_0 \quad \text{versus} \quad H_A : f = f_A. \tag{18.55}
\]

For given \( \alpha \in [0, 1] \), we wish to find a test function \( \phi \) that

\[
\text{Maximizes } E_A[\phi(X)] \text{ subject to } E_0[\phi(X)] \leq \alpha. \tag{18.56}
\]

A test function \( \psi \) has Neyman-Pearson form if for some constant \( c \in [0, \infty] \) and function \( \gamma(x) \in [0, 1] \),

\[
\psi(x) = \begin{cases} 
1 & \text{if } LR(x) > c \\
\gamma(x) & \text{if } LR(x) = c \\
0 & \text{if } LR(x) < c
\end{cases}, \tag{18.57}
\]

with the caveat that

\[
\text{if } c = \infty \text{ then } \gamma(x) = 1 \text{ for all } x. \tag{18.58}
\]

Here,

\[
LR(x) = \frac{f_A(x)}{f_0(x)} \in [0, \infty] \tag{18.59}
\]

is defined unless \( f_A(x) = f_0(x) = 0 \), in which case \( \psi \) is unspecified. Notice that \( LR \) and \( c \) are allowed to take on the value \( \infty \).

**Lemma 32. Neyman-Pearson Lemma** Any test \( \psi \) of Neyman-Pearson form (18.57,18.58) for which \( E_0[\psi(X)] = \alpha \) satisfies conditions (18.56).

So basically, the likelihood ratio test is best. One can take the \( \gamma(x) \) to be a constant, but sometimes it is convenient to have it depend on \( x \). Before getting to the proof, consider some special cases, of mainly theoretical interest.

- \( \alpha = 0 \). If there is no chance of rejecting when the null is true, then one must always accept if \( f_0(x) > 0 \), and it always makes sense to reject when \( f_0(x) = 0 \). Such actions invoke the caveat (18.58), that is, when \( f_A(x) > 0 \),

\[
\begin{cases} 
1 & \text{if } LR(x) = \infty \\
0 & \text{if } LR(x) < \infty
\end{cases} = \begin{cases} 
1 & \text{if } f_0(x) = 0 \\
0 & \text{if } f_0(x) > 0
\end{cases} \tag{18.60}
\]

- \( \alpha = 1 \). This one is silly from a practical point of view, but if you do not care about rejecting when the null is true, then you should always reject, i.e., take \( \phi(x) = 1 \).

- **Power = 1.** If you want to be sure to reject if the alternative is true, then \( \phi(x) = 1 \) when \( f_A(x) > 0 \), so take the test (18.57) with \( c = 0 \). Of course, you may not be able to achieve your desired \( \alpha \).
Proof of Lemma 32. If $\alpha = 0$, then the above discussion shows that taking $c = \infty, \gamma(x) = 1$ (18.58) is best. For $\alpha \in (0, 1)$, suppose $\psi$ satisfies (18.57) for some $c$ and $\gamma(x)$ with $E_0[\psi(X)] = \alpha$, and $\phi$ is any other test function with $E_0[\phi(X)] \leq \alpha$. Look at

$$E_A[\psi(X) - \phi(X)] - c \ E_0[\psi(X) - \phi(X)] = \int_X (\psi(x) - \phi(x)) f_A(x) dx$$

$$+ c \int_X (\psi(x) - \phi(x)) f_0(x) dx$$

$$= \int_X (\psi(x) - \phi(x)) LR(x) f_0(x) dx$$

$$- c \int_X (\psi(x) - \phi(x)) f_0(x) dx$$

$$= \int_X (\psi(x) - \phi(x)) (LR(x) - c) f_0(x) dx$$

$$\geq 0. \quad (18.61)$$

The final inequality holds because $\psi = 1$ if $LR - c > 0$, and $\psi = 0$ if $LR - c < 0$, so that the final integrand is always nonnegative. Thus

$$E_A[\psi(X) - \phi(X)] \geq c \ E_0[\psi(X) - \phi(X)]$$

$$\geq 0, \quad (18.62)$$

because $E_0[\psi(X)] = \alpha \geq E_0[\phi(X)]$. Thus $E_A[\psi(X)] \geq E_A[\phi(X)]$, i.e., any other level $\alpha$ test has lower or equal power. $\square$

There are a couple of addenda to the lemma which we will not prove here. First, for any $\alpha$, there is a test of Neyman-Pearson form. Second, if the $\phi$ in the proof is not essentially of Neyman-Pearson form, then the power of $\psi$ is strictly better than that of $\phi$.

18.4.1 Examples

If $f_0(x) > 0$ and $f_A(x) > 0$ for all $x \in \mathcal{X}$, then it is straightforward (though maybe not easy) to find the Neyman-Pearson test. It can get tricky if one or the other density is 0 at times.

Normal means. Suppose $\mu_0$ and $\mu_A$ are fixed, $\mu_A > \mu_0$, and $X \sim N(\mu, 1)$. We wish to test

$$H_0 : \mu = \mu_0 \ vs \ H_A : \mu = \mu_A \quad (18.63)$$

with $\alpha = 0.05$. Here,

$$LR(x) = \frac{e^{-\frac{1}{2} (x-\mu_A)^2}}{e^{-\frac{1}{2} (x-\mu_0)^2}} = e^{x(\mu_A-\mu_0) - \frac{1}{2}(\mu_A^2-\mu_0^2)}. \quad (18.64)$$

Because $\mu_A > \mu_0$, $LR(x)$ is strictly increasing in $x$, so $LR(x) > c$ is equivalent to $x > c^*$ for some $c^*$. For level 0.05, we know that $c^* = 1.645 + \mu_0$, so the test must reject when $LR(x) > LR(c^*)$, i.e.,

$$\psi(x) = \begin{cases} 1 & \text{if } e^{x(\mu_A-\mu_0) - \frac{1}{2}(\mu_A^2-\mu_0^2)} > c, \\ 0 & \text{if } e^{x(\mu_A-\mu_0) - \frac{1}{2}(\mu_A^2-\mu_0^2)} \leq c, \quad c = e^{1.645 + \mu_0}(\mu_A-\mu_0) - \frac{1}{2}(\mu_A^2-\mu_0^2). \end{cases} \quad (18.65)$$
We have taken the $\gamma = 0$; the probability $LR(X) = c$ is 0, so it doesn’t matter what happens then. That expression is unnecessarily complicated. In fact, to find $c$ we already simplified the test, that is

$$LR(x) > c \iff x - \mu_0 > 1.645,$$

hence

$$\psi(x) = \begin{cases} 1 & \text{if } x - \mu_0 > 1.645 \\ 0 & \text{if } x - \mu_0 \leq 1.645 \end{cases}.$$  

That is, we really do not care about $c$, as long as we have the $\psi$.

**Double Exponential versus Normal.** Now suppose $f_0$ is the Double Exponential pdf and $f_A$ is the $N(0, 1)$ pdf, and $\alpha = 0.1$. Then

$$LR(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} x^2} = \sqrt{\frac{2}{\pi}} e^{\frac{1}{2} |x|^2}.$$  

Now $LR(x) > c$ if and only if

$$|x| - \frac{1}{2} x^2 > c^* = \log(c) + \frac{1}{2} \log(\pi/2)$$

if and only if (completing the square)

$$(|x| - 1)^2 < c^{**} = -2c^* - 1 \iff ||x| - 1| < c^{***} = \sqrt{c^{**}}.$$  

We need to find the constant $c^{***}$ so that

$$P_0(||x| - 1| < c^{***}) = 0.10, \ X \sim \text{Double Exponential}.$$  

For a smallish $c^{***}$, using the Double Exponential pdf,

$$P_0(||X| - 1| < c^{***}) = P_0[-1 - c^{***} < X < -1 + c^{***} \text{ or } 1 - c^{***} < X < 1 + c^{***}] = 2 P_0[-1 - c^{***} < X < -1 + c^{***}] = e^{-\left(1-c^{***}\right)} - e^{-\left(1+c^{***}\right)}.$$  

Setting that equal to 0.10, we find $c^{***} = 0.1355$. The picture shows a horizontal line at 0.1355. The rejection region consists of the $x$’s for which the graph of $||x| - 1|$ is below the line.
The power is
\[ P_A[|X| - 1 < 0.1355] = P_A[-1 - 0.1355 < X < -1 + 0.1355 \quad or \quad 1 - 0.1355 < X < 1 + 0.1355] = 2(\Phi(1.1355) - \Phi(0.8645)) = 0.1311. \] (18.73)

Not very powerful, but at least it is larger than \( \alpha \! \)! Of course, it is not surprising that it is hard to distinguish the normal from the double exponential with just one observation.

**Uniform versus Uniform I.** Suppose \( X \sim Uniform(0, \theta) \), and the question is whether \( \theta = 1 \) or 2. We could then test
\[ H_0 : \theta = 1 \quad versus \quad H_A : \theta = 2. \] (18.74)

The likelihood ratio is
\[ LR(x) = \frac{f_A(x)}{f_0(x)} = \frac{\frac{1}{\theta} I[0 < x < 2]}{\frac{1}{\theta} I[0 < x < 1]} = \begin{cases} \frac{1}{2} & \text{if } 0 < x < 1 \\ \infty & \text{if } 1 \leq x < 2. \end{cases} \] (18.75)

No matter what, you would reject the null if \( 1 \leq x < 2 \), because it is impossible to observe an \( x \) in that region under the \( Uniform(0, 1) \).

First try \( \alpha = 0 \). Usually, that would mean never reject, so power would be 0 as well, but here it is not that bad. We invoke (18.58), that is, take \( c = \infty \) and \( \gamma(x) = 1 \):
\[ \psi(x) = \begin{cases} 1 & \text{if } LR(x) = \infty \\ 0 & \text{if } LR(x) < \infty \end{cases} = \begin{cases} 1 & \text{if } 1 \leq x < 2 \\ 0 & \text{if } 0 < x < 1. \end{cases} \] (18.76)
18.4. The Neyman-Pearson Lemma

Then

\[ \alpha = P[1 \leq U(0,1) < 2] = 0 \quad \text{and} \quad \text{Power} = P[1 \leq U(0,2) < 2] = \frac{1}{2}. \]  

(18.77)

What if \( \alpha = 0.1 \)? Then the Neyman-Pearson test would take \( c = \frac{1}{2} \):

\[ \psi(x) = \begin{cases} 1 & \text{if } LR(x) > \frac{1}{2} \\ \gamma(x) & \text{if } LR(x) = \frac{1}{2} \\ 0 & \text{if } LR(x) < \frac{1}{2} \end{cases} = \begin{cases} 1 & \text{if } 1 \leq x < 2 \\ \gamma(x) & \text{if } 0 < x < 1 \end{cases}, \]

(18.78)

because \( LR \) cannot be less than \( \frac{1}{2} \). Notice that

\[ E_0[\psi(X)] = E_0[\gamma(X)] = \int_0^1 \gamma(x) dx, \]

(18.79)

so that any \( \gamma \) than integrates to \( \alpha \) works. Some examples:

\[ \gamma(x) = \begin{cases} 0.1 & \text{if } 0 < x < 0.1 \\ 1 & \text{if } 0.9 < x < 1 \\ 0.2 x & \text{if } 1 < x < 2 \end{cases}. \]

(18.80)

No matter which you choose, the power is the same:

\[ \text{Power} = E_A[\psi(X)] = \frac{1}{2} \int_0^1 \gamma(x) dx + \frac{1}{2} \int_1^2 dx = \frac{1}{2} \alpha + \frac{1}{2} = 0.55. \]

(18.81)

**Uniform versus Uniform II.** Now switch the null and alternative in (18.74), keeping \( X \sim \text{Uniform}(0,\theta) \):

\[ H_0 : \theta = 2 \quad \text{versus} \quad H_A : \theta = 1. \]

(18.82)

Then the likelihood ratio is

\[ LR(x) = \frac{f_A(x)}{f_0(x)} = \frac{I[0 < x < 1]}{\frac{1}{2} I[0 < x < 2]} = \begin{cases} 2 & \text{if } 0 < x < 1 \\ 0 & \text{if } 1 \leq x < 2 \end{cases}. \]

(18.83)

For level \( \alpha = 0.1 \), we have to take \( c = 2 \):

\[ \psi(x) = \begin{cases} \gamma(x) & \text{if } LR(x) = 2 \\ 0 & \text{if } LR(x) < 2 \end{cases} = \begin{cases} \gamma(x) & \text{if } 0 < x < 1 \\ 0 & \text{if } 1 \leq x < 2 \end{cases}. \]

(18.84)

Then any \( \gamma(x) \) with

\[ \frac{1}{2} \int_0^1 \gamma(x) dx = \alpha \implies \int_0^1 \gamma(x) dx = 0.2 \]

(18.85)

will work. And that is the power,

\[ \text{Power} = \int_0^1 \gamma(x) dx = 0.2. \]

(18.86)
18.4.2 Bayesian approach

The Neyman-Pearson approach is all about action: Either accept the null or reject it. As in the courts, it does not suggest the degree to which the null is plausible or not. By contrast, the Bayes approach produces the probabilities the null and alternative are true, given the data and a prior distribution. That is, we start with the prior:

\[ P[H_0] (= P[H_0 \text{ is true}]) = \pi_0, \quad P[H_A] = \pi_A, \]  
(18.87)

where \( \pi_0 + \pi_A = 1 \). Where do these probabilities come from? Presumably, from a reasoned consideration of all that is known prior to seeing the data. Or, one may try to be fair and take \( \pi_0 = \pi_A = \frac{1}{2} \). The densities are then the conditional distributions of \( X \) given the hypotheses are true:

\[ X \mid H_0 \sim f_0 \quad \text{and} \quad X \mid H_A \sim f_A. \]  
(18.88)

Bayes Theorem gives the posterior probabilities:

\[
P[H_0 \mid X = x] = \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_A f_A(x)}, \quad P[H_A \mid X = x] = \frac{\pi_A f_A(x)}{\pi_0 f_0(x) + \pi_A f_A(x)}. \]
(18.89)

Dividing numerators and denominators by \( f_0(x) \) (assuming it is not 0), we have

\[
P[H_0 \mid X = x] = \frac{\pi_0}{\pi_0 + \pi_A LR(x)}, \quad P[H_A \mid X = x] = \frac{\pi_A LR(x)}{\pi_0 + \pi_A LR(x)}, \]
(18.90)

showing that these posteriors depend on the data only through the likelihood ratio. That is, the posterior does not violate the likelihood principle. Hypothesis tests do violate the likelihood principle: Whether they reject depends on the \( c \), which is calculated from \( f_0 \), not the likelihood.

Odds are actually more convenient here, where the odds of an event \( A \) are

\[
Odds[A] = \frac{P[A]}{1 - P[A]}, \quad \text{hence} \quad P[A] = \frac{Odds[A]}{1 + Odds[A]}.
\]  
(18.91)

The prior odds in favor of \( H_A \) is then \( \pi_a / \pi_0 \), and the posterior odds are

\[
Odds[H_A \mid X = x] = \frac{P[H_A \mid X = x]}{P[H_0 \mid X = x]} = \frac{\pi_A LR(x)}{\pi_0} = Odds[H_A] \times LR(x).
\]  
(18.92)

That is,

\[
\text{Posterior odds} = (\text{Prior odds}) \times (\text{Likelihood ratio}),
\]  
(18.93)

which neatly separates the contribution to the posterior of the prior and the data. If a decision is needed, one would choose a cutoff point \( k \), say, and reject the null if the posterior odds exceed \( k \), and maybe randomize if they equal \( k \). But this test is the same as the Neyman-Pearson test (18.57) with \( k = c\pi_0 / \pi_A \). The difference is that \( k \) would not be chosen to achieve a certain level, but rather on decision-theoretic grounds, that is, depending on the relative costs of making a Type I or Type II error.
For example, take the example in (18.41), where \( X \sim \text{Binomial}(4, p) \), and the hypotheses are

\[
H_0 : p = \frac{1}{2} \quad \text{versus} \quad H_A : p = \frac{3}{5}.
\]  

(18.94)

If the prior odds are even, i.e., \( \pi_0 = \pi_A = \frac{1}{2} \), so that the prior odds are 1, then the posterior odds are equal to the likelihood ratio, giving the following:

| \( x \) | Odds \([H_A | X = x] = LR(x)\) | \( P[H_A | X = x] \) |
|---|---|---|
| 0  | 0.4096 | 0.2906 |
| 1  | 0.6144 | 0.3806 |
| 2  | 0.9216 | 0.4796 |
| 3  | 1.3824 | 0.5803 |
| 4  | 2.0736 | 0.6746 |

(18.95)

Thus if you see \( X = 2 \) heads, the posterior probability that \( p = \frac{3}{5} \) is about 48%, and if \( X = 4 \), it is about 67%. None of these probabilities is overwhelmingly in favor of one hypothesis or the other.

18.5 Uniformly most powerful tests

Simple versus simple is too simple. Some testing problems are not so simple, and yet do have a best test. Here is the formal definition.

**Definition 34.** The test function \( \psi \) is a **uniformly most powerful** (UMP) level \( \alpha \) test for testing \( H_0 : \theta \in \Theta_0 \) versus \( H_A : \theta \in \Theta_A \) if it is level \( \alpha \) and

\[
E_{\theta}[\psi(X)] \geq E_{\theta}[\phi(X)] \quad \text{for all} \quad \theta \in \Theta_A
\]

for any other level \( \alpha \) test \( \phi \).

Often, one-sided tests do have a UMP test, while two-sided tests do not. For example, suppose \( X \sim N(\mu, 1) \), and we test whether \( \mu = 0 \). In a one-sided testing problem, the alternative is one of \( \mu > 0 \) or \( \mu < 0 \), say

\[
H_0 : \mu = 0 \quad \text{versus} \quad H_A^{(1)} : \mu > 0.
\]  

(18.97)

The corresponding two-sided testing problem is

\[
H_0 : \mu = 0 \quad \text{versus} \quad H_A^{(2)} : \mu \neq 0.
\]  

(18.98)

Level \( \alpha = 0.05 \) tests for these are, respectively,

\[
\phi^{(1)}(x) = \begin{cases} 
1 & \text{if} \quad x > 1.645 \\
0 & \text{if} \quad x \leq 1.645
\end{cases}
\quad \text{and} \quad \phi^{(2)}(x) = \begin{cases} 
1 & \text{if} \quad |x| > 1.96 \\
0 & \text{if} \quad |x| \leq 1.96
\end{cases}.
\]  

(18.99)

Their powers are

\[
E_\mu[\phi^{(1)}(X)] = P[N(\mu, 1) > 1.645] = \Phi(\mu - 1.645) \quad \text{and}
\]

\[
E_\mu[\phi^{(2)}(X)] = P[|N(\mu, 1)| > 1.96] = \Phi(\mu - 1.96) + \Phi(-\mu - 1.96).
\]  

(18.100)

See the graph:
For the one-sided problem, the power is good for $\mu > 0$, but bad (below $\alpha$) for $\mu < 0$. But the alternative is just $\mu > 0$, so it does not matter what $\phi^{(1)}$ does when $\mu < 0$. For the two-sided test, the power is fairly good on both sides of $\mu = 0$, but it is not quite as good as the one-sided test when $\mu > 0$. The other line in the graph is the one-sided test for alternative $\mu < 0$, which mirrors $\phi^{(1)}$, rejecting when $x < -1.645$.

The following are true, and to be proved:

- For the one-sided problem (18.97), the test $\phi^{(1)}$ is the UMP level $\alpha = 0.05$ test.

- For the two-sided problem (18.98), there is no UMP level $\alpha$ test. Test $\phi^{(1)}$ is better on one side ($\mu > 0$), and the other one-sided test is better on the other side. None of the three tests is always best.

We start with the null being simple and the alternative being composite (which just means not simple). The way to prove a test is UMP level $\alpha$ is to show that it is level $\alpha$, and that it is of Newman-Pearson form for each simple versus simple subproblem derived from the big problem. That is, suppose we are testing

\[ H_0 : \theta = \theta_0 \text{ versus } H_A : \theta \in \Theta_A. \]  

A simple versus simple subproblem takes a specific value from the alternative, so that for a given $\theta_A \in \Theta_A$, we consider

\[ H_0 : \theta = \theta_0 \text{ versus } H_A^{(\theta_A)} : \theta = \theta_A. \]  

**Theorem 14.** Suppose that for testing problem (18.101), $\psi$ satisfies

\[ E_{\theta_0}[\psi(X)] = \alpha \]  

(18.103)
and that for each $\theta \in \Theta$,

$$\psi(x) = \begin{cases} 
1 & \text{if } LR(x; \theta) > c(\theta), \\
\gamma(x) & \text{if } LR(x; \theta) = c(\theta), \\
0 & \text{if } LR(x; \theta) < c(\theta),
\end{cases}$$

for some constant $c(\theta)$, (18.104)

where

$$LR(x; \theta) = \frac{f_{\theta}(x)}{f_{\theta_0}(x)}.$$ (18.105)

Then $\psi$ is a UMP level $\alpha$ test for (18.101).

**Proof.** Suppose $\phi$ is another level $\alpha$ test. Then for the subproblem (18.102), $\psi$ has at least as high power, i.e.,

$$E_{\theta_A}[\psi(X)] \geq E_{\theta_A}[\phi(X)].$$ (18.106)

But that inequality is true for any $\theta \in \Theta$, hence $\psi$ is UMP level $\alpha$. $\square$

The difficulty is to find a test $\psi$ which is Neyman-Pearson for all $\theta_A$. Consider the example with $X \sim N(\mu, 1)$ and hypotheses

$$H_0 : \mu = 0 \quad \text{versus} \quad H_A^{(1)} : \mu > 0,$$ (18.107)

and take $\alpha = 0.05$. For fixed $\mu > 0$, the Neyman-Pearson test is

$$\psi(x) = \begin{cases} 
1 & \text{if } LR(x; \mu_A) > c(\mu), \\
\gamma(x) & \text{if } LR(x; \mu_A) = c(\mu), \\
0 & \text{if } LR(x; \mu_A) < c(\mu),
\end{cases}$$

where

$$LR(x; \mu_A) = \frac{e^{-\frac{1}{2}((x-\mu)^2-x^2)}}{e^{-\frac{1}{2}((x-\mu_0)^2-x^2)}}.$$ (18.108)

The last step is valid because we know $\mu_A > 0$. That messy constant is chosen so that the level is 0.05, which we know must be

$$\log(c(\mu_A)) + \frac{1}{2} \mu_A^2/\mu_A = 1.645.$$ (18.109)

The key point is that (18.109) is true for any $\mu_A > 0$, that is,

$$\psi(x) = \begin{cases} 
1 & \text{if } x > 1.645 \\
0 & \text{if } x \leq 1.645
\end{cases}$$ (18.110)

is true for any $\mu_A$. Thus $\psi$ is indeed UMP. Note that the constant $c(\mu)$ is different for each $\mu_A$, but the test $\psi$ is the same.

Why is there no UMP test for the two-sided problem,

$$H_0 : \mu = 0 \quad \text{versus} \quad H_A^{(1)} : \mu \neq 0?$$ (18.111)
The best test at alternative $\mu_A > 0$ is (18.110), but the best test at alternative $\mu_A < 0$ is

$$\psi(x) = \begin{cases} 1 & \text{if } LR(x; \mu_A) > c(\mu_A) \\ 0 & \text{if } LR(x; \mu_A) \leq c(\mu_A) \end{cases}$$

$$= \begin{cases} 1 & \text{if } x\mu_A > \log(c(\mu_A)) + \frac{1}{2} \mu_A^2 \\ 0 & \text{if } x\mu_A \leq \log(c(\mu_A)) + \frac{1}{2} \mu_A^2 \end{cases}$$

$$= \begin{cases} 1 & \text{if } x < (\log(c(\mu_A)) + \frac{1}{2} \mu_A^2)/\mu_A \\ 0 & \text{if } x \geq (\log(c(\mu_A)) + \frac{1}{2} \mu_A^2)/\mu_A \end{cases}$$

$$= \begin{cases} 1 & \text{if } x < -1.645 \\ 0 & \text{if } x \geq -1.645 \end{cases}$$

(18.112)

which is different than test (18.110). That is, there is no test that is best at both positive and negative values of $\mu_A$, so there is no UMP test.

### 18.5.1 One-sided exponential family testing problems

The normal example about can be extended to general exponential families. The key to the existence of a UMP test is that the $LR(x; \theta_A)$ is increasing in some function no matter what the alternative. That is, suppose $X_1, \ldots, X_n$ are iid with a one-dimensional exponential family density

$$f(x \mid \theta) = a(x)e^{\theta \Sigma(x_i) - n\psi(\theta)}.$$ (18.113)

A one-sided testing problem is

$$H_0 : \theta = \theta_0 \text{ versus } H_A : \theta > \theta_0.$$ (18.114)

Then for fixed alternative $\theta_A > \theta_0$,

$$LR(x; \theta_A) = \frac{f(x \mid \theta_A)}{f(x \mid \theta_0)} = e^{(\theta_A - \theta_0)\Sigma(x_i) - n(\psi(\theta_A) - \psi(\theta_0))}.$$ (18.115)

Then the best test at the alternative $\theta_A$ is

$$\psi(x) = \begin{cases} 1 & \text{if } LR(x; \theta_A) > c(\theta_A) \\ 1 & \text{if } LR(x; \theta_A) = c(\theta_A) \\ 0 & \text{if } LR(x; \theta_A) \leq c(\theta_A) \end{cases}$$

$$= \begin{cases} 1 & \text{if } (\theta_A - \theta_0)\Sigma T(x_i) > \log(c(\theta_A)) + n(\psi(\theta_A) - \psi(\theta_0)) \\ \gamma & \text{if } (\theta_A - \theta_0)\Sigma T(x_i) = \log(c(\theta_A)) + n(\psi(\theta_A) - \psi(\theta_0)) \\ 0 & \text{if } (\theta_A - \theta_0)\Sigma T(x_i) \leq \log(c(\theta_A)) + n(\psi(\theta_A) - \psi(\theta_0)) \end{cases}$$

$$= \begin{cases} 1 & \text{if } \sum T(x_i) > (\log(c(\theta_A)) + n(\psi(\theta_A) - \psi(\theta_0)))/(\theta_A - \theta_0) \\ \gamma & \text{if } \sum T(x_i) = (\log(c(\theta_A)) + n(\psi(\theta_A) - \psi(\theta_0)))/(\theta_A - \theta_0) \\ 0 & \text{if } \sum T(x_i) \leq (\log(c(\theta_A)) + n(\psi(\theta_A) - \psi(\theta_0)))/(\theta_A - \theta_0) \end{cases}$$

(18.116)

Then $c$ and $\gamma$ are chosen to give the right level, but they are the same for any alternative $\theta_A > \theta_0$. Thus the test (18.116) is UMP level $\alpha$.

If the alternative were $\theta < \theta_0$, then the same reasoning would work, but the inequalities would switch. For a two-sided alternative, there would not be a UMP test.
18.5.2 Monotone likelihood ratio

A generalization of exponential families, which guarantee UMP tests, are families with monotone likelihood ratio. Non-exponential family examples include the non-central $\chi^2$ and $F$ distributions.

**Definition 35.** A family of densities $f(x \mid \theta), \theta \in \Theta \subset \mathbb{R}$ has monotone likelihood ratio (MLR) with respect to parameter $\theta$ and statistic $T(x)$ if for any $\theta' < \theta$,

$$\frac{f(x \mid \theta)}{f(x \mid \theta')}$$

is a function of just $T(x)$, and is nondecreasing in $T(x)$. If the ratio is strictly increasing in $T(x)$, then the family has strict monotone likelihood ratio.

Note in particular that this $T(x)$ is a sufficient statistic. It is fairly easy to see that one-dimensional exponential families have MLR. The general idea of MLR is that in some sense, as $\theta$ gets bigger, $T(X)$ gets bigger. Two lemmas follow, without proofs.

**Lemma 33.** If the family $f(x \mid \theta)$ has MLR with respect to $\theta$ and $T(x)$, then for any nondecreasing function $g(w)$,

$$E_{\theta}[g(T(X))] \text{ is nondecreasing in } \theta.$$  

(18.118)

If the family has strict MLR, and $g$ is strictly increasing, then the expected value in (18.118) is strictly increasing.

**Lemma 34.** Suppose the family $f(x \mid \theta)$ has MLR with respect to $\theta$ and $T(x)$, and we are testing

$$H_0 : \theta = \theta_0 \text{ versus } H_A : \theta > \theta_0$$

(18.119)

for some level $\alpha$. Then the test

$$\psi(x) = \begin{cases} 1 & \text{if } \sum T(x_i) > c \\ \gamma & \text{if } \sum T(x_i) = c \\ 0 & \text{if } \sum T(x_i) < c \end{cases}$$

(18.120)

where $c$ and $\gamma$ are chosen to achieve level $\alpha$, is UMP level $\alpha$.

In the situation in Lemma 34, the power function $E_{\theta}[\psi(X)]$ is increasing, because of Lemma 33, since $\psi$ is nondecreasing in $\theta$. In fact, MLR also can be used to show that the test (18.120) is UMP level $\alpha$ for testing

$$H_0 : \theta \leq \theta_0 \text{ versus } H_A : \theta > \theta_0$$

(18.121)

18.6 Some additional criteria

It is quite special to actually have a UMP test. Similar to the estimation case, where we may restrict to unbiased estimators or equivariant estimators, in testing we may wish to find UMP tests among those satisfying certain other criteria. Here are a few, pertaining to the general testing problem

$$H_0 : \theta \in \Theta_0 \text{ versus } H_A : \theta \in \Theta_A.$$  

(18.122)
1. **Unbiased.** A test $\psi$ is **unbiased** if

$$E_{\Theta_A}[\psi(X)] \geq E_{\Theta_0}[\psi(X)] \quad \text{for any } \Theta_0 \in \Theta_0 \text{ and } \Theta_A \in \Theta_A. \quad (18.123)$$

This means that you are more likely to reject when you should than when you shouldn’t. Note that in the two-sided Normal mean example, the one-sided test (18.98) are not unbiased, because their power falls below $\alpha$ on one side of 0. The two-sided test (18.99) is unbiased, and is in fact the UMP unbiased level $\alpha = 0.05$ test.

2. **Similar.** A test $\psi$ is **similar** if it equals the level for all null values, i.e.,

$$E_{\Theta}[\psi(X)] = \alpha \quad \text{for all } \Theta \in \Theta_0. \quad (18.124)$$

For example, the $t$-test in (18.16) and the sign test (18.17) are both similar for testing $\mu = \mu_0$ in the Normal case where $\sigma^2$ is unknown, as in (18.15). That is because the distributions of those two statistics do not depend on $\sigma^2$ when $\mu = \mu_0$. The $t$-test turns out to be the UMP similar level $\alpha$ test.

3. **Invariant.** A testing problem is invariant under a group $G$ if *both* the null and alternative models are invariant under the group. A test $\psi$ is invariant if $\psi(h_g(\underline{x})) = \psi(\underline{x})$ for all $g \in G$, where $h_g$ is the group’s action on $\mathcal{X}$. (Recall Chapter 13.5.)
The previous chapter concentrates on fairly simple testing situations, primarily one-dimensional. When the parameter spaces are more complex, it is difficult if not impossible to find optimal procedures. But many good procedures still are based on the likelihood ratio. For the general testing problem

\[ H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_A : \theta \in \Theta_A, \]  

consider the likelihood ratio between two fixed parameters,

\[ \frac{f(x | \theta_A)}{f(x | \theta_0)} \quad \text{for some} \quad \theta_0 \in \Theta_0 \quad \text{and} \quad \theta_A \in \Theta_A. \]  

How to extend this ratio to all \( \theta \)'s? Here are three popular approaches:

- **Score**: Differentiate. If the null hypothesis is simple, then the ratio, or actually its log, can be expanded about \( \theta_A = \theta_0 \):

  \[ \log \left( \frac{f(x | \theta_A)}{f(x | \theta_0)} \right) = I(\hat{\theta}_A; x) - I(\hat{\theta}_0; x) \approx (\hat{\theta}_A - \theta_0)'S(\theta_0; x). \]  

  That test statistic is approximately the best statistic for alternative \( \theta_A \) when \( \theta_A \) is very close to \( \theta_0 \).

- **Maximize the likelihoods**: This approaches picks the best value of \( \theta \) from each model. That is, \( \hat{\theta}_A \) is the MLE from the alternative model, and \( \hat{\theta}_0 \) is the MLE from the null model. Then the test statistic is

  \[ 2 \log \left( \frac{\sup_{\theta_A \in \Theta_A} f(x | \theta_A)}{\sup_{\theta_0 \in \Theta_0} f(x | \theta_0)} \right) = 2 \log \left( \frac{f(x | \hat{\theta}_A)}{f(x | \hat{\theta}_0)} \right) = 2(I(\hat{\theta}_A; x) - I(\hat{\theta}_0; x)). \]  

  The "2" is just a detail to make the null distribution easier to approximate by a \( \chi^2 \).
Chapter 19. Likelihood-based testing

• **Average: The Bayes factor.** Here, one has probability densities $\rho_A$ over $\Theta_A$ and $\rho_0$ over $\Theta_0$, and looks at the ratio of the averages of the likelihoods:

$$\text{Bayes Factor} = \frac{\int_{\Theta_A} f(x | \theta_A) \rho_A(\theta_A) d\theta_A}{\int_{\Theta_0} f(x | \theta_0) \rho_0(\theta_0) d\theta_0}. \quad (19.5)$$

This can take the place of $LR$ in (18.93):

$$\text{Posterior odds} = (\text{Prior odds}) \times (\text{Bayes Factor}). \quad (19.6)$$

We will look at the first two approaches in more detail in the next two sections.

One nice property of these test statistics is that, under certain conditions, it is easy to approximate the null distributions if $n$ is large.

### 19.1 Score tests

#### 19.1.1 One-sided

Suppose $X_1, \ldots, X_n$ are iid with density $f(x_i | \theta)$, where $\theta \in \mathbb{R}$. We start with the simple versus one-sided testing problem,

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_A : \theta > \theta_0. \quad (19.7)$$

For fixed $\theta_A > \theta_0$, the test which rejects the null when $(\theta_A - \theta_0)S_n(\theta_0; x) > c$ is then the same that rejects when $S_n(\theta_0; x) > c^*$. ($S_n$ is the score for all $n$ observations, and $S_n(\theta; x) = \sum_{i=1}^n S_1(\theta; x_i)$, where $S_1(\theta; x_i)$ is the score for just $x_i$.) Under the null hypothesis, we know that

$$E_{\theta_0}[S_1(\theta_0; X_i)] = 0 \quad \text{and} \quad \text{Var}_{\theta_0}[S_1(\theta_0; X_i)] = I_1(\theta_0), \quad (19.8)$$

hence by the central limit theorem,

$$T_n(x) = \frac{S_n(\theta_0; X)}{\sqrt{n I_1(\theta_0)}} \rightarrow^D N(0, 1) \quad \text{under the null hypothesis.} \quad (19.9)$$

Then the one-sided score test

Rejects the null hypothesis when $T_n(x) > z_{\alpha}$, \quad (19.10)

which is approximately level $\alpha$.

**Example.** Suppose the $X_i$’s are iid Cauchy$(\theta, 1)$, so that

$$f(x_i | \theta) = \frac{1}{\pi} \frac{1}{1 + (x_i - \theta)^2}. \quad (19.11)$$

We wish to test

$$H_0 : \theta = 0 \quad \text{versus} \quad H_A : \theta > 0, \quad (19.12)$$

so that $\theta_0 = 0$. The score at $\theta = 0$ and information in one observation are, respectively,

$$S_1(0; x_i) = \frac{2x_i}{1 + x_i^2} \quad \text{and} \quad I_1(0) = \frac{1}{2}. \quad (19.13)$$
Then the test statistic is

\[ T_n(x) = \frac{S_n(0; x)}{\sqrt{n I_1(0)}} = \frac{\sum_{i=1}^{n} \frac{2x_i}{1+x_i^2}}{\sqrt{n/2}} = \frac{2\sqrt{2}}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i}{1+x_i^2}, \quad (19.14) \]

and for an approximate 0.05 level, the cutoff point is \( z_{0.05} = 1.645 \). If the information is difficult to calculate, which it is not in this example, then you can use the observed information at \( \theta_0 \) instead.

\[ \square \]

19.1.2 Many-sided

Now consider a more general problem, where \( \theta \in \Theta \subset \mathbb{R}^K \), and the null, \( \theta_0 \), is in the interior of \( \Theta \). The testing problem is

\[ H_0: \theta = \theta_0 \quad \text{versus} \quad H_A: \theta \in \Theta - \{ \theta_0 \}. \quad (19.15) \]

For a fixed value of the alternative \( \theta_A \), under null we have that

\[ E_{\theta_0}[(\theta_A - \theta_0)' S_n(\theta_0; X)] = 0_K \quad \text{and} \quad \text{Var}_{\theta_0}[(\theta_A - \theta_0)' S_n(\theta_0; X)] = n (\theta_A - \theta_0)' I_1(\theta_0)(\theta_A - \theta_0), \quad (19.16) \]

because the covariance matrix of the score is the Fisher Information matrix. Then the score statistic for \( \theta_A \) is

\[ T_n(x; \theta_A) = \frac{(\theta_A - \theta_0)' S_n(\theta_0; X)}{\sqrt{n (\theta_A - \theta_0)' I_1(\theta_0)(\theta_A - \theta_0)}}. \quad (19.17) \]

For any particular \( \theta_A \), that statistic is approximately \( N(0, 1) \) under the null. But we do not want to have to specify an alternative. Because our test statistic is supposed to measure the distance from the null, we could take the \( \theta_A \) that maximizes the \( T_n(x; \theta_A) \), that is, the statistic is

\[ T_n(x) = \max_{\theta_A \neq \theta_0} T_n(x; \theta_A). \quad (19.18) \]

To find that maximum, we can simplify by taking

\[ a = \frac{I_1^{1/2}(\theta_0)(\theta_A - \theta_0)}{\sqrt{(\theta_A - \theta_0)' I_1(\theta_0)(\theta_A - \theta_0)}}, \quad \text{so that} \quad \|a\| = 1. \quad (19.19) \]

We are assuming that \( I_1 \) is invertible, and mean by \( I_1^{1/2} \) a matrix square root of \( I_1 \). We can then write

\[ T_n(x) = \max_a \left\{ a' b \right\}, \quad \text{where} \quad b = \frac{1}{\sqrt{n}} I_1^{-1/2}(\theta_0) S_n(\theta_0; X). \quad (19.20) \]

The Cauchy-Schwarz inequality says that

\[ (a' b)^2 \leq \|a\|^2 \|b\|^2 = \|b\|^2, \quad (19.21) \]
because $\|a\| = 1$. Can that upper bound be achieved? Yes, at least if $b \neq 0$, by taking $a = b/\|b\|$ (which does have length 1), so that

$$
(a'b)^2 = ((b/\|b\|)'b)^2 = (b'b)^2 / \|b\|^2 = \|b\|^4 / \|b\|^2 = \|b\|^2.
$$

(19.22)

The bound (19.21) is true for $b = 0$, too, of course. Going back to (19.20), then, we obtain

$$
T_n^2(x) = \frac{1}{n} \mathbb{S}_n(\theta_0; X)'I_1^{-1}(\theta_0)\mathbb{S}_n(\theta_0; X).
$$

(19.23)

This $T_n^2$ is the official score statistic for this problem. Note that the dependence on $\theta_A$ is gone.

We still need a cutoff point. The multivariate central limit theorem says that under the null hypothesis,

$$
\frac{1}{\sqrt{n}} \mathbb{S}_n(\theta_0; X) \approx N(0, I_K),
$$

(19.24)

hence

$$
\frac{1}{\sqrt{n}} I_1^{-1/2}(\theta_0)\mathbb{S}_n(\theta_0; X) \approx N(0, I_K),
$$

(19.25)

where this $I_K$ is the identity matrix. Thus $T_n^2$ is the sum of squares of approximately iid $N(0, 1)$'s, so that

$$
T_n^2(x) \approx \chi^2_K,
$$

(19.26)

and the approximate level $\alpha$ score test rejects the null hypothesis if

$$
T_n^2(x) > \chi^2_{K, \alpha}.
$$

(19.27)

Example. Suppose $X \sim Multinomial_2(n, (p_1, p_2, p_3))$, and we wish to test that the probabilities are equal:

$$
H_0 : p_1 = p_2 = \frac{1}{3} \quad \text{versus} \quad H_A : (p_1, p_2) \neq \left(\frac{1}{3}, \frac{1}{3}\right).
$$

(19.28)

We’ve left out $p_3$ because it is a function of the other two. Leaving it in will cause the information matrix to be noninvertible.

The loglikelihood in one Multinoulli $Z = (Z_1, Z_2, Z_3)$ is

$$
l(p_1, p_2; Z) = z_1 \log(p_1) + z_2 \log(p_2) + z_3 \log(1 - p_1 - p_2).
$$

(19.29)

The score at the null is then

$$
\mathbb{S}_1(1/3, 1/3; Z) = \begin{pmatrix}
\frac{\partial l(p_1, p_2; Z)}{\partial p_1} \\
\frac{\partial l(p_1, p_2; Z)}{\partial p_2}
\end{pmatrix}_{p_1=1/3, p_2=1/3} = \begin{pmatrix}
z_1/p_1 - z_3/(1 - p_1 - p_2) \\
z_2/p_2 - z_3/(1 - p_1 - p_2)
\end{pmatrix}_{p_1=1/3, p_2=1/3} = 3 \begin{pmatrix}
z_1 - z_3 \\
z_2 - z_3
\end{pmatrix}.
$$

(19.30)
The information is the covariance matrix of the score. To find that, we first have under the null that

$$\text{Cov}_0[Z] = \begin{pmatrix} \frac{2}{9} & -\frac{1}{9} & -\frac{1}{9} \\ -\frac{1}{9} & \frac{2}{9} & -\frac{1}{9} \\ -\frac{1}{9} & -\frac{1}{9} & \frac{2}{9} \end{pmatrix} = \frac{1}{9} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}. \quad (19.31)$$

Because

$$S_1(1/3, 1/3; Z) = 3 \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} Z,$$ \quad (19.32)

the information is

$$I_1(1/3, 1/3) = \text{Cov}[S_1(1/3, 1/3; Z)]$$ \quad (19.33)

$$= \frac{3^2}{9} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 6 & 3 \\ 3 & 6 \end{pmatrix}. \quad (19.34)$$

The score for all $n$ observations is the same as for one, but with $X_i$’s instead of the $Z_i$’s. Thus

$$T^2_n(x) = \frac{1}{n} (3 \begin{pmatrix} X_1 - X_3 \\ X_2 - X_3 \end{pmatrix})' I_1(1/3, 1/3)^{-1} (3 \begin{pmatrix} X_1 - X_3 \\ X_2 - X_3 \end{pmatrix})'$$

$$= \frac{1}{n} \left( \begin{pmatrix} X_1 - X_3 \\ X_2 - X_3 \end{pmatrix} \right)' \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} X_1 - X_3 \\ X_2 - X_3 \end{pmatrix}'$$

$$= \frac{1}{n} (2(X_1 - X_3)^2 + 2(X_2 - X_3)^2 - 2(X_1 - X_3)(X_2 - X_3))$$

$$= \frac{1}{n} ((X_1 - X_3)^2 + (X_2 - X_3)^2 + (X_1 - X_2)^2). \quad (19.35)$$

The cutoff point is $\chi^2_{2, \alpha}$, because there are $K = 2$ parameters. The statistic looks reasonable, because the $X_i$’s would tend to be different if their $p_i$’s were. Also, it may not look like it, but this $T^2_n$ is the same as the Pearson $\chi^2$ statistic for these hypotheses,

$$X^2 = \sum_{i=1}^{3} \frac{(x_i - n/3)^2}{n/3} = \sum_{i=1}^{3} \frac{(\text{Obs}_i - \text{Exp}_i)^2}{\text{Exp}_i}, \quad (19.36)$$

where $\text{Obs}_i$ is the observed count $x_i$, and $\text{Exp}_i$ is the expected count under the null, which here is $n/3$ for each $i$.

### 19.2 The maximum likelihood ratio test

Even more general than the score test is the maximum likelihood ratio test (MLRT) (or sometimes just LRT), which uses the likelihood ratio statistic, but substitutes the
MLE for the parameter value in each density. The statistic is
\[
\Lambda(x) = \sup_{\theta_A \in \Theta_A} f(x | \theta_A) \sup_{\theta_0 \in \Theta_0} f(x | \theta_0) = \frac{f(x | \hat{\theta}_A)}{f(x | \hat{\theta}_0)},
\] (19.37)
where \(\hat{\theta}_0\) is the MLE of \(\theta\) under the null model, and \(\hat{\theta}_A\) is the MLE of \(\theta\) under the alternative model. Notice that in the simple versus simple situation, \(\Lambda(x) = LR(x)\). This statistic may or may not lead to a good test. In many situations it is quite reasonable, in the same way that the MLE is often a good estimator. Also, under appropriate conditions, it is easy to find the cutoff point to obtain the approximate level \(\alpha\):

Under \(H_0\), \(2 \log(\Lambda(X)) \rightarrow^D \chi^2_v\), \(v = \text{dimension}(\Theta_A) - \text{dimension}(\Theta_0)\). (19.38)

First we will show some examples, then formalize the above result.

19.2.1 Examples

Example 1. Suppose \(X_1, \ldots, X_n\) are iid \(N(\mu, \sigma^2)\), and we wish to test whether \(\mu = 0\), with \(\sigma^2\) unknown:

\[H_0 : \mu = 0, \sigma^2 > 0 \text{ versus } H_A : \mu \neq 0, \sigma^2 > 0.\] (19.39)

Here, \(\theta = (\mu, \sigma^2)\), so we need to find the MLE under the two models.

- **Null.** Here, \(\Theta_0 = \{(0, \sigma^2) \mid \sigma^2 > 0\}\). Thus the MLE of \(\mu\) is \(\hat{\mu}_0 = 0\), since it is the only possibility. For \(\sigma^2\), we then maximize

\[
\frac{1}{\sigma^n} e^{-\sum x_i^2/(2\sigma^2)},
\] (19.40)

which by the usual calculations yields

\[
\hat{\sigma}^2_A = \frac{\sum x_i^2}{n}, \text{ hence } \hat{\theta}_0 = (0, \sum x_i^2 / n).
\] (19.41)

- **Alternative.** Now \(\Theta_A = \{(\mu, \sigma^2) \mid \mu \neq 0, \sigma^2 > 0\}\), which from long ago we know yields MLE’s

\[
\hat{\theta}_0 = (\bar{x}, s^2), \text{ where } s^2 = \hat{\sigma}_0^2 = \frac{\sum (x_i - \bar{x})^2}{n}.
\] (19.42)

Notice that not only is the MLE of \(\mu\) different in the two models, but so is the MLE of \(\sigma^2\). (Which should be reasonable, because if you know the mean is 0, you do not have to use \(\bar{x}\).) Next, stick those estimates into the likelihood ratio, and see what
19.2. The maximum likelihood ratio test

happens (the $\sqrt{2\pi}$'s cancel):

$$\Lambda(x) = \frac{f(x \mid \bar{x}, s^2)}{f(x \mid 0, \sum x_i^2 / n)}$$

$$= \frac{1}{s^2 e^{-\Sigma(x_i - \bar{x})^2 / (2s^2)}} \frac{1}{(\Sigma x_i^2 / n)^{n/2}} e^{-\Sigma x_i^2 / (2\Sigma x_i^2 / n)}$$

$$= \left( \frac{\sum x_i^2 / n}{s^2} \right)^{n/2} e^{-n/2} e^{-n/2}$$

$$= \left( \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)^{n/2} .$$  (19.43)

Thus the MLRT

Rejects $H_0$ when $\Lambda(x) > c$  (19.44)

for some cutoff $c$. To find $c$, we need the distribution of $\Lambda(X)$ under the null, which does not look obvious. But we can start to fool around. There are a couple of ways to go. First,

$$\Lambda(x) > c \iff \frac{\sum (x_i - \bar{x})^2}{\sum x_i^2} < c^* = c^{-2/n}$$

$$\iff \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2 + n\bar{x}^2} < c^*$$

$$\iff \frac{\sum (x_i - \bar{x})^2 / \sigma^2}{\sum (x_i - \bar{x})^2 / \sigma^2 + n\bar{x}^2 / \sigma^2} < c^* .$$  (19.45)

We know that $\sum (X_i - \bar{X})^2$ and $\bar{X}$ are independent, and

$$\frac{\sum (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{and} \quad \frac{n\bar{X}^2}{\sigma^2} \sim \chi^2_1 \text{ (since } \mu = 0).$$  (19.46)

But $\chi^2_1$ is $\Gamma(n/2, 1/2)$. Thus, under the null hypothesis,

$$\frac{\sum (X_i - \bar{X})^2 / \sigma^2}{\sum (X_i - \bar{X})^2 / \sigma^2 + n\bar{X}^2 / \sigma^2} \sim \text{Beta} \left( \frac{n - 1}{2}, \frac{1}{2} \right) ,$$  (19.47)

and the $c^*$ can be found from the Beta distribution.

Or, if you do not have a Beta table handy,

$$\Lambda(x) > c \iff \frac{\sum (x_i - \bar{x})^2 + n\bar{x}^2}{\sum (x_i - \bar{x})^2} > c^* = c^{2/n}$$

$$\iff \frac{n\bar{x}^2}{\sum (x_i - \bar{x})^2} > c^{**} = c^* - 1$$

$$\iff \frac{(\sqrt{n}\bar{x})^2}{\sum (x_i - \bar{x})^2 / (n - 1)} > c^{***} = (n - 1)c^{**}$$

$$\iff |T_n(x)| > c^{****} = \sqrt{c^{***}} .$$  (19.48)
where $T_n$ is the $t$-statistic,

$$T_n(x) = \frac{\sqrt{n}x}{\sqrt{\sum(x_i - \bar{x})^2/(n-1)}}.$$  

Thus the cutoff is $c^{**} = t_{n-1, a/2}$. Which is to say, the MLRT in this case is the usual two-sided $t$-test.

**Example 2.** In (19.28), we tested $p = (p_1, p_2, p_3) = (1/3, 1/3, 1/3)$ for $X \sim \text{Multinomial}_3(n, p)$. The likelihood is

$$L(p_1, p_2, p_3; x) = p_1^{x_1} p_2^{x_2} p_3^{x_3},$$

ignoring the factorial terms. To find the MLRT, we again need to find two MLE’s. Because the null hypothesis is simple, the MLE for the null is just $\hat{\theta}_0 = (1/3, 1/3, 1/3)$. For the alternative, the $p$ is unrestricted, in which case the MLE is $\hat{\theta}_A = (x_1/n, x_2/n, x_3/n)$. Thus

$$\Lambda(x) = \frac{p_{A1}^x p_{A2}^x p_{A3}^x}{p_{01}^x p_{02}^x p_{03}^x} = \frac{(x_1/n)^{x_1} (x_2/n)^{x_2} (x_3/n)^{x_3}}{(1/3)^{x_1} (1/3)^{x_2} (1/3)^{x_3}} = \left(\frac{3}{n}\right)^n x_1^{x_1} x_2^{x_2} x_3^{x_3}. (19.51)$$

Notice that this statistic is quite a bit different from the score statistic in (19.35). What is the distribution of $\Lambda(x)$ under the null? We will address that question in the next subsection.

### 19.2.2 Asymptotic null distribution

Similar to the way that, under conditions, the MLE is asymptotically (multivariate) normal, the $2 \log(\Lambda(x))$ is asymptotically $\chi^2$ under the null. We will not present the proof of the general result, but sketch it when the null is simple. We have $X_1, \ldots, X_n$ iid, each with density $f^*(x_i | \theta), \theta \in \Theta \subset \mathbb{R}^K$, and make the assumptions in Section 17.1 used for likelihood estimation. Consider testing

$$H_0 : \theta = \theta_0 \text{ versus } H_A : \theta \in \Theta - \{\theta_0\}$$

for fixed $\theta_0 \in \Theta$. One of the assumptions is that $\Theta$ is an open set, so that $\theta_0$ is in the interior of $\Theta$. In particular, this assumption rules out one-sided tests. The MLE under the null is thus the fixed $\hat{\theta}_0 = \theta_0$, and the MLE under the alternative, $\hat{\theta}_A$, is the usual MLE.

From (19.37), with

$$l_n(\theta; x) = \log(f(x | \theta)) = \sum_{i=1}^n \log(f^*(x_i | \theta)),$$

we have that

$$2 \log(\Lambda(x)) = 2(l_n(\hat{\theta}_A; x) - l_n(\theta_0; x)).$$

(19.54)
Expand $l_n(\hat{\theta}_i; x)$ around $\hat{\theta}_A$:  

$$l_n(\theta_0; x) \approx l_n(\hat{\theta}_A; x) + (\theta_0 - \hat{\theta}_A)'S_n(\hat{\theta}_A; x) + \frac{1}{2} (\theta_0 - \hat{\theta}_A)'H_n(\hat{\theta}_A; x)(\theta_0 - \hat{\theta}_A),$$  

(19.55)

where $H_n$ is the second derivative matrix,

$$H_n(\hat{\theta}; x) = \begin{pmatrix}
\frac{\partial^2 l_n(\hat{\theta}; x)}{\partial \theta_1^2} & \frac{\partial^2 l_n(\hat{\theta}; x)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 l_n(\hat{\theta}; x)}{\partial \theta_1 \partial \theta_K} \\
\frac{\partial^2 l_n(\hat{\theta}; x)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 l_n(\hat{\theta}; x)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 l_n(\hat{\theta}; x)}{\partial \theta_2 \partial \theta_K} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 l_n(\hat{\theta}; x)}{\partial \theta_K \partial \theta_1} & \frac{\partial^2 l_n(\hat{\theta}; x)}{\partial \theta_K \partial \theta_2} & \cdots & \frac{\partial^2 l_n(\hat{\theta}; x)}{\partial \theta_K^2}
\end{pmatrix}. \quad (19.56)$$

Then $-H_n/n$ is the observed Fisher Information matrix.

Looking at (19.55), we have that $S_n(\hat{\theta}_A; x) = 0_K$, because the score is the first derivative of the log likelihood, so that by (19.54),

$$2\log(\Lambda(x)) \approx -(\theta_0 - \hat{\theta}_A)'H_n(\hat{\theta}_A; x)(\theta_0 - \hat{\theta}_A) = -(\sqrt{n}(\theta_0 - \hat{\theta}_A))'(H_n(\hat{\theta}_A; x)/n)(\sqrt{n}(\theta_0 - \hat{\theta}_A)). \quad (19.57)$$

Now suppose $H_0$ is true, so that $\hat{\theta}_0$ is the true value of the parameter. Then as in (17.74),

$$\sqrt{n} (\hat{\theta}_A - \theta_0) \longrightarrow^D N(0_K, I^{-1}(\theta_0)), \quad (19.58)$$

and as in (17.55),

$$\frac{1}{n} H_n(\hat{\theta}_A; x) \longrightarrow^P I_1(\hat{\theta}_0). \quad (19.59)$$

Thus

$$2\log(\Lambda(x)) \longrightarrow^D N(0_K, I^{-1}(\theta_0))'I_1(\hat{\theta}_0)N(0_K, I^{-1}(\theta_0)) \sim \chi^2_K. \quad (19.60)$$

Now recall the multinomial example with $\Lambda(x)$ in (19.51). The $\hat{\theta} = (p_1, p_2)$, even though there are three $p_i$’s, because they sum to 0. Thus $K = 2$. We can check whether the parameterization is ok by looking at the Fisher Information, because by assumption (17.73) it must be invertible. If we take $\hat{\theta} = (p_1, p_2, p_3)$, then the information will not be invertible.

The $\Theta = \{(p_1, p_2) \mid 0 < p_1, 0 < p_2, p_1 + p_2 < 1\}$, which is indeed open, and the null $\theta_0 = (1/3, 1/3)$ is in the interior. Thus, under the null,

$$2\log(\Lambda(x)) = 2n \log(3/n) + 2 \sum x_i \log(x_i) \longrightarrow^D \chi^2_2. \quad (19.61)$$

A perhaps more familiar expression is

$$2\log(\Lambda(x)) = 2 \sum x_i \log(x_i/(n/3)) = 2 \sum \text{Obs}_i \log(\text{Obs}_i/\text{Exp}_i), \quad (19.62)$$

where as before, $\text{Obs}_i$ is the observed count $x_i$, and $\text{Exp}_i$ is the expected count under the null. This statistic is called the likelihood ratio $\chi^2$, not unreasonably, which is a competitor to the Pearson $\chi^2$ in (19.36).
Composite null

Again with \( \Theta \subseteq \mathbb{R}^K \), where \( \Theta \) is open, we consider the null hypothesis that sets part of \( \theta \) to zero. That is, partition

\[
\theta = \begin{pmatrix} \theta^{(1)} \\ \theta^{(2)} \end{pmatrix}, \quad \theta^{(1)} \text{ is } K_1 \times 1, \quad \theta^{(2)} \text{ is } K_2 \times 1, \quad K_1 + K_2 = K. \tag{19.63}
\]

The problem is to test

\[
H_0 : \theta^{(1)} = 0_{K_1} \text{ versus } H_A : \theta^{(1)} \neq 0_{K_1}, \tag{19.64}
\]

with \( \theta^{(2)} \) unspecified. More precisely,

\[
H_0 : \theta \in \Theta_0 = \{ \theta \in \Theta \mid \theta^{(1)} = 0_{K_1} \} \text{ versus } H_A : \theta \in \Theta_A = \Theta - \Theta_0. \tag{19.65}
\]

The parameters in \( \theta^{(2)} \) are called nuisance parameters, because they are not of primary interest, but still need to be dealt with. Without them, we would have a nice simple null. The main results follows.

**Theorem 15.** If the Cramér assumptions in Section 17.1 hold, then under the null, the MLRT statistic \( \Lambda \) for problem (19.65) has

\[
2 \log(\Lambda(X)) \xrightarrow{D} \chi^2_{K_1}. \tag{19.66}
\]

This theorem takes care of the simple null case as well, where \( K_2 = 0 \).

**Example 1.** Setting some \( \theta_i \)'s to zero may seem to be an overly restrictive type of null, but many testing problems can be reparameterized into that form. For example, suppose \( X_1, \ldots, X_n \) are iid \( N(\mu_X, \sigma^2) \), and \( Y_1, \ldots, Y_n \) are iid \( N(\mu_Y, \sigma^2) \), and the \( X_i \)'s and \( Y_i \)'s are independent. We wish to test \( \mu_X = \mu_Y \), with \( \sigma^2 \) unknown. Then the hypotheses are

\[
H_0 : \mu_X = \mu_Y, \sigma^2 > 0 \text{ versus } H_A : \mu_X \neq \mu_Y, \sigma^2 > 0. \tag{19.67}
\]

To put these in the form (19.65), take a one-to-one reparameterizations

\[
\theta = \begin{pmatrix} \mu_X - \mu_Y \\ \mu_X + \mu_Y \\ \sigma^2 \end{pmatrix}; \quad \theta^{(1)} = \mu_X - \mu_Y \text{ and } \theta^{(2)} = \begin{pmatrix} \mu_X + \mu_Y \\ \sigma^2 \end{pmatrix}. \tag{19.68}
\]

Then

\[
\Theta_0 = \{0\} \times \mathbb{R} \times (0, \infty) \text{ and } \Theta = \mathbb{R}^2 \times (0, \infty). \tag{19.69}
\]

Here, \( K_1 = 1 \), and there are \( K_2 = 2 \) nuisance parameters, \( \mu_X + \mu_Y \) and \( \sigma^2 \). Thus the asymptotic \( \chi^2 \) has 1 degree of freedom.

**Example 2.** Consider \( X \sim \text{Multinomial}_4(n, p) \), where \( p = (p_1, p_2, p_3, p_4)' \), arranged as a \( 2 \times 2 \) contingency table, where the factors are performance on Homework and Exams:

<table>
<thead>
<tr>
<th>HW ↓; Exams →</th>
<th>lo</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>lo</td>
<td>( X_1 )</td>
<td>( X_2 )</td>
</tr>
<tr>
<td>hi</td>
<td>( X_3 )</td>
<td>( X_4 )</td>
</tr>
</tbody>
</table>
19.2. The maximum likelihood ratio test

The null hypothesis we want to test is that performance on Exams and HW are independent. One approach is to find a confidence interval for the log odds ratio \( \gamma = \log(p_1 p_4 / (p_2 p_3)) \), and see whether 0 is in the interval. Here we will find the MLRT. Let \( q_1 = P[HW = lo] \) and \( r_1 = P[Exams = lo] \), etc., giving the table

<table>
<thead>
<tr>
<th></th>
<th>lo</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>lo</td>
<td>( p_1 )</td>
<td>( p_2 )</td>
</tr>
<tr>
<td>hi</td>
<td>( p_3 )</td>
<td>( p_4 )</td>
</tr>
<tr>
<td></td>
<td>( q_1 = p_1 + p_2 )</td>
<td>( q_2 = p_3 + p_4 )</td>
</tr>
</tbody>
</table>

The null hypothesis is then

\[
H_0 : p_1 = q_1 r_1, \ p_2 = q_1 r_2, \ p_3 = q_2 r_1, \ p_4 = q_2 r_2. \tag{19.72}
\]

But we have to reparameterize so that the null sets something to 0. First, we know that we can drop \( p_4 \), say. Then consider

\[
(p_1, p_2, p_3) \longleftrightarrow (p_1, q_1, r_1) \longleftrightarrow (p_1 - q_1 r_1, q_1, r_1) \equiv \hat{\theta}. \tag{19.73}
\]

Check that these correspondences are indeed one-to-one. (E.g., \( p_2 = q_1 - p_1 \), etc.) Also, check that \( p_1 - q_1 r_1 = 0 \) implies independence as in (19.72). Then

\[
\hat{\theta}^{(1)} = p_1 - q_1 r_1 \quad \text{and} \quad \hat{\theta}^{(2)} = \left( \begin{array}{c} q_1 \\ r_1 \end{array} \right). \tag{19.74}
\]

So, in particular, \( K_1 = 1 \).

Now for the MLRT. The likelihood is

\[
L(p; \hat{x}) = p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4}. \tag{19.75}
\]

Under the null as in (19.72),

\[
L(p; \hat{x}) = (q_1 r_1)^{x_1} (q_1 (1 - r_1))^{x_2} ((1 - q_1) r_1)^{x_3} ((1 - q_1) (1 - r_1))^{x_4} = q_1^{x_1 + x_2} (1 - q_1)^{x_3 + x_4} r_1^{x_1 + x_3} (1 - r_1)^{x_2 + x_4}. \tag{19.76}
\]

Note that that looks like the two binomial likelihoods multiplied together, so we know what the MLE’s are: \( \hat{q}_1 = (x_1 + x_2) / n \) and \( \hat{r}_1 = (x_1 + x_3) / n \), and we also have automatically that \( \hat{\theta}^{(1)} = p_1 - q_1 r_1 = 0 \).

Under the alternative, we know that \( \hat{p}_i = x_i / n \) for each \( i \). Then from those we can find the MLE of \( \hat{\theta} \), but actually we do not need to, because all we need is the maximum likelihood. We now have

\[
\Lambda(x) = \frac{(x_1 / n)^{x_1} (x_2 / n)^{x_2} (x_3 / n)^{x_3} (x_4 / n)^{x_4}}{(\hat{q}_1 \hat{r}_1)^{\hat{x}_1} (\hat{q}_2 \hat{r}_2)^{\hat{x}_2} (\hat{q}_1 \hat{r}_1)^{\hat{x}_3} (\hat{q}_2 \hat{r}_2)^{\hat{x}_4}}. \tag{19.77}
\]

Shifting the \( n \)'s to the denominator, we have

\[
2 \log(\Lambda(x)) = 2 \left( x_1 \log \left( \frac{x_1}{n \hat{q}_1 \hat{r}_1} \right) + x_2 \log \left( \frac{x_2}{n \hat{q}_1 \hat{r}_2} \right) \right. \\
\left. + x_3 \log \left( \frac{x_3}{n \hat{q}_2 \hat{r}_1} \right) + x_4 \log \left( \frac{x_4}{n \hat{q}_2 \hat{r}_2} \right) \right) = \sum \text{Obs}_i \log \left( \frac{\text{Obs}_i}{\text{Exp}_i} \right), \tag{19.78}
\]

where \( \text{Obs}_i \) is the observed frequency and \( \text{Exp}_i \) is the expected frequency under the null hypothesis.
where $\text{Exp}_i$ is the MLE of the expected value of $X_i$ under the null. Under the null, this statistic is asymptotically $\chi^2_1$.

One lesson is that it is fine to use different parameterizations for the two hypotheses when finding the respective MLE’s. Expressing the hypotheses in terms of the $\theta$ is mainly to make sure the theorem can be used, and to find the $K_1$. 
Chapter 20

Nonparametric Tests and Randomization

There are two steps in testing a pair of hypotheses: Finding a test statistic with good power, and finding its distribution under the null hypothesis. The previous chapter showed that good test statistics tend to arise from the likelihood ratio. Finding the null distribution depended on either getting lucky by being able to express as a random variable with a familiar distribution, or by using the asymptotic normality or \( \chi^2 \)-ity if certain assumptions held. This chapter takes a more expansive approach in not necessarily being tied to parametric models, or at least wishing to be robust against departures from the model, and in dealing with test statistics that may not have an easy null distribution, even approximately.

To illustrate, recall the example (18.15) with \( X_1, \ldots, X_n \) iid \( N(\mu, \sigma^2) \), where we test whether \( \mu = \mu_0 \). Two possible statistics were the (absolute) \( t \)-statistic and sign statistic, respectively

\[
T_n(x) = \frac{|x - \mu_0|}{s/\sqrt{n}} \quad \text{and} \quad S_n(x) = \left| \frac{\# \{ x_i < \mu_0 \}}{n} - \frac{1}{2} \right|. \tag{20.1}
\]

It turns out that the \( t \)-test has better power than the sign test. (Both statistics are scale-invariant, in the sense that multiplying the \( X_i - \mu_0 \)'s by the same non-zero constant does not change the statistic. Then it can be shown that the two-sample \( t \)-test is the UMP level \( \alpha \) invariant test. But we won’t go into that, as interesting as it is.)

There are several ways our view of this testing problem can be expanded:

- **Robustness.** The distribution of the \( X_i \)'s may not be exactly normal, but something close, or at least normal but with the possibility of outliers. It still would make sense to test whether the mean is zero, but what happens to the two statistics? From the central limit theorem and Slutsky, we know that as long as the \( X_i \)'s are iid and have finite variance, \( T_n(x) \to^D N(0,1) \), so that at least an approximate level \( \alpha \) test is to reject when \( T_n > 1.96 \). The sign test is also valid, as long as \( \mu_0 \) is the mean and the median, since \( Y_n = \# \{ x_i < \mu_0 \} \) would be \( \text{Bin}(n, 1/2) \) anytime the \( X_i \)'s are iid (and continuous) with median \( \mu_0 \). Under those conditions, the same cutoff point would work whether normality held or not. Whether the power of the tests would be adequate is another question.

- **Nonparametric.** Rather than assume normality or near-normality, one can instead change the problem to a totally nonparametric one. For example, suppose
$X_1, \ldots, X_n$ are iid, $X_i \sim F$, where

$$F \in \mathcal{F} = \{F \mid F \text{ is a continuous distribution function with finite variance that is symmetric about the median } \eta \}.$$  \hfill (20.2)

That symmetry means $P[X - \eta > c] = P[X - \eta < -c]$, i.e., $1 - F(c + \eta) = F(-c + \eta)$ for any $c$. That set of distribution functions is called non-parametric because it is not indexed by a finite-dimensional parameter $\theta$. That doesn’t mean there aren’t parameters, e.g., means and variances, etc. Notice that those $F$’s include the normal, double exponential and logistic, but not Cauchy (because the Cauchy does not have finite variance) nor exponential, which is not symmetric. Then the testing problem would be

$$H_0 : F \in \mathcal{F} \text{ with } \eta = \eta_0 \text{ versus } H_A : F \in \mathcal{F} \text{ with } \eta \neq \eta_0 \hfill (20.3)$$

for some fixed $\eta_0$. Notice the $F$ is a nuisance parameter (or nuisance non-parameter, if you wish). This problem is testing whether the median is a specific value, but by the conditions on $F$, the median and the mean are the same. If we removed the symmetry condition, then we would not be testing whether the mean is $\eta_0$. Sign and rank tests are often used for such nonparametric hypotheses, because their distribution does not depend on $F$.

**Randomization.** We may wish to use a statistic whose distribution under the null is difficult to find or even approximate well, whether because of the statistic itself or because of the lack of specificity in the null. Instead of finding the distribution of the test statistic, we look at the statistic as the data is randomly tweaked in a way consistent with the null hypothesis. The randomization distribution of the statistic $T(x)$ is the result of this tweaking.

In the next few sections, we will look at some specific testing problems and their nonparametric counterparts, showing how to find the randomization distribution of any test statistic, and exhibiting some popular signs and rank tests.

### 20.1 Testing the median

As discussed above, the nonparametric analogs to hypotheses on the mean are hypotheses on the median. Thus to repeat, we assume that $X_1, \ldots, X_n$ are iid, $X_i \sim F \in \mathcal{F}$, where $\mathcal{F}$ is the set of continuous distribution functions that have finite variance and are symmetric about their median $\eta$, as in (20.2). We test

$$H_0 : F \in \mathcal{F} \text{ with } \eta = 0 \text{ versus } H_A : F \in \mathcal{F} \text{ with } \eta \neq 0. \hfill (20.4)$$

(What follows is easy to extend to tests of $\eta = \eta_0$: just subtract $\eta_0$ from all the $X_i$’s.) The two-sided $t$-test is valid for this problem, although it is not similar, that is, the null distribution of $T$ depends on exactly which $F$ holds. The sign test depends on the signs of the data, $\text{Sign}(x_1), \ldots, \text{Sign}(x_n)$, where for any $z \in \mathbb{R}$,

$$\text{Sign}(z) = \begin{cases} +1 & \text{if } z > 0 \\ 0 & \text{if } z = 0 \\ -1 & \text{if } z < 0. \end{cases} \hfill (20.5)$$
The two-sided sign statistic is then

\[ S(x) = \left| \sum_{i=1}^{n} \text{Sign}(x_i) \right|. \] (20.6)

If none of the \( x_i \)'s is exactly 0, then this statistic is the same as that in (18.17). The exact distribution is the same as \(|2Y - n|\) where \( Y \sim \text{Binomial}(n, 1/2)\). It is also easy to approximate, either by using the approximation to the binomial, or noting that under the null,

\[ E_0[\text{Sign}(X_i)] = 0 \quad \text{and} \quad \text{Var}_0[\text{Sign}(X_i)] = E_0[\text{Sign}(X_i)^2] = 1. \] (20.7)

Then an approximate level \( \alpha \) test rejects the null when

\[ \frac{S(x)}{\sqrt{n}} > z_{\alpha}/2. \] (20.8)

As mentioned in the introduction to this chapter, the \( t \)-test is also valid, the statistic begin approximately \( N(0, 1) \) for large \( n \). Which has better power? It depends on what the \( F \) is. If it is Normal, the \( t \)-test is best. If it is Double Exponential, the sign test is best. In fact, considering Pitman efficiency, which we will not go into (STAT 575 does) but is analogous to efficiency in estimation, the relative Pitman efficiency of the \( t \)-test to the sign test is the same as the asymptotic efficiency of the mean to the median in estimating a location parameter.

20.1.1 A randomization distribution

One could also think of other tests, for example, the two-sided “median” test, which rejects \( H_0 \) when

\[ M(x) = |\text{Median}\{x_1, \ldots, x_n\}| > c. \] (20.9)

The problem here is that without knowing the \( F \), we cannot even use the normal approximation to the distribution, because that involves \( 1/4f^2(0) \). There are ways to estimate the density, but that adds an extra layer of complication. Instead, we can use the randomization distribution of \( M(x) \). The idea is randomly change the \( x_i \)'s is a way consistent with the null hypothesis, and see what happens to the test statistic. More precisely, find a group \( G \) under which the distribution of \( X \) is unchanged for any \( F \) in the null. That is, where \( h_g \) is the action, if the null is true,

\[ h_g(X) = D X \quad \text{for any} \quad g \in G. \] (20.10)

For this testing problem, the group is the set of \( \pm 1 \)'s:

\[ G = \{(\epsilon_1, \ldots, \epsilon_n) \mid \epsilon_i = \pm 1\}, \] (20.11)

and the action is to multiply the \( x_i \)'s by the \( \epsilon_i \)'s:

\[ h_{(\epsilon_1, \ldots, \epsilon_n)}(x_1, \ldots, x_n) = (\epsilon_1x_1, \ldots, \epsilon_nx_n). \] (20.12)

This is just a fancy way of changing the signs on some of the \( x_i \)'s. Then the requirement (20.10) is true since \( X_i \) and \(-X_i\) have the same distribution under the null, i.e., when the median is 0. Notice it is not true under the alternative.
Now switch from a fixed $g$ and random $\ X$ to the fixed data $\ x$ and random $\ G$. That is, suppose $G$ is uniformly distributed over $\ G$, that is,

$$P[G = g] = \frac{1}{\#G}, \ g \in G. \quad (20.13)$$

For our $G$, there are $2^n$ elements, and what we have is that $\epsilon_1, \ldots, \epsilon_n$ are iid Uniform $\{-1, +1\}$, i.e.,

$$P[\epsilon_i = +1] = P[\epsilon_i = -1] = \frac{1}{2}. \quad (20.14)$$

Look at the random variable $M(h_G(x))$ defined by

$$M(h_G(x)) = M(\epsilon_1 x_1, \ldots, \epsilon_n x_n), \quad (20.15)$$

where the $\epsilon_i$’s are random. The distribution of $M(h_G(x))$ is the randomization distribution of $M(x)$. The cutoff point we use is the $c(x)$ that gives the desired $\alpha$:

$$P[M(h_G(x)) > c(x) \mid X = x] \leq (\text{but } \approx) \alpha. \quad (20.16)$$

(Or we could randomize where $M = c$ to get exact $\alpha$.) The conditional “$X = x$” is there just to emphasize that it is not the $x$ that is random, but the $\epsilon_i$’s.

To illustrate, suppose $n = 5$ and the data are

$$x_1 = 24.3, \ x_2 = 10.9, \ x_3 = -4.6, \ x_4 = 14.6, \ x_5 = 10.0, \ \Rightarrow M(\bar{x}) = 10.9. \quad (20.17)$$

Take $\alpha = 0.25$ (which is a bit big, but $n$ is only 5).
Here are the possible randomizations of the data, and the values of $M(h_G(x))$, and the $\text{Mean}$.

<table>
<thead>
<tr>
<th>$\epsilon_1x_1$</th>
<th>$\epsilon_2x_2$</th>
<th>$\epsilon_3x_3$</th>
<th>$\epsilon_4x_4$</th>
<th>$\epsilon_5x_5$</th>
<th>$M(h_G(x))$</th>
<th>$\text{Mean}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.3</td>
<td>10.9</td>
<td>-4.6</td>
<td>14.6</td>
<td>10</td>
<td>10.9</td>
<td>11.04</td>
</tr>
<tr>
<td>24.3</td>
<td>10.9</td>
<td>-4.6</td>
<td>14.6</td>
<td>-10</td>
<td>10.9</td>
<td>7.04</td>
</tr>
<tr>
<td>24.3</td>
<td>10.9</td>
<td>-4.6</td>
<td>-14.6</td>
<td>10</td>
<td>10</td>
<td>5.2</td>
</tr>
<tr>
<td>24.3</td>
<td>10.9</td>
<td>4.6</td>
<td>-14.6</td>
<td>-10</td>
<td>4.6</td>
<td>1.2</td>
</tr>
<tr>
<td>24.3</td>
<td>10.9</td>
<td>4.6</td>
<td>14.6</td>
<td>10</td>
<td>10.9</td>
<td>12.88</td>
</tr>
<tr>
<td>24.3</td>
<td>10.9</td>
<td>4.6</td>
<td>14.6</td>
<td>-10</td>
<td>10.9</td>
<td>8.88</td>
</tr>
<tr>
<td>24.3</td>
<td>10.9</td>
<td>4.6</td>
<td>-14.6</td>
<td>10</td>
<td>10</td>
<td>7.04</td>
</tr>
<tr>
<td>24.3</td>
<td>10.9</td>
<td>4.6</td>
<td>-14.6</td>
<td>-10</td>
<td>4.6</td>
<td>3.04</td>
</tr>
<tr>
<td>24.3</td>
<td>-10.9</td>
<td>-4.6</td>
<td>14.6</td>
<td>10</td>
<td>10</td>
<td>6.68</td>
</tr>
<tr>
<td>24.3</td>
<td>-10.9</td>
<td>-4.6</td>
<td>14.6</td>
<td>-10</td>
<td>4.6</td>
<td>2.68</td>
</tr>
<tr>
<td>24.3</td>
<td>-10.9</td>
<td>-4.6</td>
<td>-14.6</td>
<td>10</td>
<td>4.6</td>
<td>0.84</td>
</tr>
<tr>
<td>24.3</td>
<td>-10.9</td>
<td>-4.6</td>
<td>-14.6</td>
<td>-10</td>
<td>10</td>
<td>3.16</td>
</tr>
<tr>
<td>24.3</td>
<td>-10.9</td>
<td>4.6</td>
<td>14.6</td>
<td>10</td>
<td>10</td>
<td>8.52</td>
</tr>
<tr>
<td>24.3</td>
<td>-10.9</td>
<td>4.6</td>
<td>14.6</td>
<td>-10</td>
<td>4.6</td>
<td>4.52</td>
</tr>
<tr>
<td>24.3</td>
<td>-10.9</td>
<td>4.6</td>
<td>-14.6</td>
<td>10</td>
<td>4.6</td>
<td>2.68</td>
</tr>
<tr>
<td>24.3</td>
<td>-10.9</td>
<td>4.6</td>
<td>-14.6</td>
<td>-10</td>
<td>10</td>
<td>1.32</td>
</tr>
<tr>
<td>-24.3</td>
<td>10.9</td>
<td>-4.6</td>
<td>14.6</td>
<td>10</td>
<td>10</td>
<td>1.32</td>
</tr>
<tr>
<td>-24.3</td>
<td>10.9</td>
<td>-4.6</td>
<td>14.6</td>
<td>-10</td>
<td>4.6</td>
<td>2.68</td>
</tr>
<tr>
<td>-24.3</td>
<td>10.9</td>
<td>-4.6</td>
<td>-14.6</td>
<td>10</td>
<td>4.6</td>
<td>4.52</td>
</tr>
<tr>
<td>-24.3</td>
<td>10.9</td>
<td>-4.6</td>
<td>-14.6</td>
<td>-10</td>
<td>10</td>
<td>8.52</td>
</tr>
<tr>
<td>-24.3</td>
<td>10.9</td>
<td>4.6</td>
<td>14.6</td>
<td>10</td>
<td>10</td>
<td>3.16</td>
</tr>
<tr>
<td>-24.3</td>
<td>10.9</td>
<td>4.6</td>
<td>14.6</td>
<td>-10</td>
<td>4.6</td>
<td>0.84</td>
</tr>
<tr>
<td>-24.3</td>
<td>10.9</td>
<td>4.6</td>
<td>-14.6</td>
<td>10</td>
<td>4.6</td>
<td>2.68</td>
</tr>
<tr>
<td>-24.3</td>
<td>10.9</td>
<td>4.6</td>
<td>-14.6</td>
<td>-10</td>
<td>10</td>
<td>6.68</td>
</tr>
<tr>
<td>-24.3</td>
<td>-10.9</td>
<td>-4.6</td>
<td>14.6</td>
<td>10</td>
<td>4.6</td>
<td>3.04</td>
</tr>
<tr>
<td>-24.3</td>
<td>-10.9</td>
<td>-4.6</td>
<td>14.6</td>
<td>-10</td>
<td>10</td>
<td>7.04</td>
</tr>
<tr>
<td>-24.3</td>
<td>-10.9</td>
<td>-4.6</td>
<td>-14.6</td>
<td>10</td>
<td>10.9</td>
<td>8.88</td>
</tr>
<tr>
<td>-24.3</td>
<td>-10.9</td>
<td>-4.6</td>
<td>-14.6</td>
<td>-10</td>
<td>10.9</td>
<td>12.88</td>
</tr>
<tr>
<td>-24.3</td>
<td>-10.9</td>
<td>4.6</td>
<td>14.6</td>
<td>10</td>
<td>4.6</td>
<td>1.2</td>
</tr>
<tr>
<td>-24.3</td>
<td>-10.9</td>
<td>4.6</td>
<td>14.6</td>
<td>-10</td>
<td>10</td>
<td>5.2</td>
</tr>
<tr>
<td>-24.3</td>
<td>-10.9</td>
<td>4.6</td>
<td>-14.6</td>
<td>10</td>
<td>10.9</td>
<td>7.04</td>
</tr>
<tr>
<td>-24.3</td>
<td>-10.9</td>
<td>4.6</td>
<td>-14.6</td>
<td>-10</td>
<td>10.9</td>
<td>11.04</td>
</tr>
</tbody>
</table>

The randomization distribution of $M$ can then be gathered from the column of $M(h_G(x))$'s:

\[
P[M(h_G(x)) = 4.6 \mid X = \overline{x}] = \frac{12}{32} = \frac{3}{8},
\]
\[
P[M(h_G(x)) = 10 \mid X = \overline{x}] = \frac{12}{32} = \frac{3}{8},
\]
\[
P[M(h_G(x)) = 10.9 \mid X = \overline{x}] = \frac{8}{32} = \frac{1}{4}.
\]

To obtain $\alpha = 0.25$, we can take the cutoff point somewhere between 10 and 10.9, say 10.5. Notice that this cutoff point depends on the data – different data would yield a different randomization distribution. The test then rejects the null if

\[M(\overline{x}) > c(\overline{x}) = 10.5.\]
The actual $M(\bar{x}) = 10.9$, so in this case we do reject.

We could also obtain the randomization distribution of $|\bar{x}|$, which is given in the last column of (20.18). Here are the values in order:

\[
\begin{array}{cccccccccccc}
0.84 & 0.84 & 1.20 & 1.20 & 1.32 & 1.32 & 2.68 & 2.68 & 2.68 & 2.68 \\
3.04 & 3.04 & 3.16 & 3.16 & 4.52 & 4.52 & 5.20 & 5.20 & 6.68 & 6.68 \\
12.88 & 12.88
\end{array}
\]

If we choose the cutoff as 8, say, then $8/32 = 0.25$ of those values exceed 8. That is, we can reject the null when $|\bar{x}| > 8$. From the data, $\bar{x} = 11.04$, so we again reject.

When $n$ is large, then it may be tedious to find the randomization distribution exactly, because there are $2^n$ elements in $G$. But we can easily use computer simulations.

**Example.** Each of $n = 16$ tires was subject to measurement of tread wear by two methods, one based on weight loss and one on groove wear. See [http://lib.stat.cmu.edu/DASL/Datafiles/differencetestdat.html](http://lib.stat.cmu.edu/DASL/Datafiles/differencetestdat.html). Thus the data are $(x_1, y_1), \ldots, (x_n, y_n)$, with the $x$ and $y$’s representing the two measurements. We assume the tires (i.e., $(x_i, y_i)$’s) are iid, and the null hypothesis is that the two measurement methods have the same distribution. (We are not assuming $X_i$ is independent of $Y_i$.) Under the null, $Z_i = X_i - Y_i$ has a distribution symmetric about 0, so we can test the null using the statistic $M(z) = |\text{Median}(z_1, \ldots, z_n)|$ again. Here are the data:

\[
\begin{array}{ccc}
X & Y & Z \\
1 & 45.9 & 35.7 & 10.2 \\
2 & 41.9 & 39.2 & 2.7 \\
3 & 37.5 & 31.1 & 6.4 \\
4 & 33.4 & 28.1 & 5.3 \\
5 & 31 & 24 & 7 \\
6 & 30.5 & 28.7 & 1.8 \\
7 & 30.9 & 25.9 & 5 \\
8 & 31.9 & 23.3 & 8.6 \\
9 & 30.4 & 23.1 & 7.3 \\
10 & 27.3 & 23.7 & 3.6 \\
11 & 20.4 & 20.9 & 0.5 \\
12 & 24.5 & 16.1 & 8.4 \\
13 & 20.9 & 19.9 & 1 \\
14 & 18.9 & 15.2 & 3.7 \\
15 & 13.7 & 11.5 & 2.2 \\
16 & 11.4 & 11.2 & 0.2
\end{array}
\]

Then $M(\bar{z}) = 4.35$. Is that significant? Rather than find the exact randomization distribution, I found 10000 samples from it, that is, 10000 times the $z_i$’s were multiplied by random $\pm1$’s, and the resulting $M(h_G(z))$’s were calculated. A histogram of the results:
There is not much beyond the observed $M(z) = 4.35$. We can count how many are greater than or equal to the observed to be 36. That is,

$$P[M(h_G(z)) \geq M(z)] = P[M(h_G(z)) \geq 4.35] \approx \frac{36}{10000} = 0.0036. \quad (20.21)$$

That is the randomization $p$-value. It is less than 0.05 by quite a bit. Thus we can easily reject the null hypothesis for $\alpha = 0.05$. (That is, the cutoff point for $\alpha = 0.05$ would be smaller than 4.35.)

### 20.2 Randomization test formalities

The advantage of randomization tests is that the exact distribution of $M(h_G(z))$ is, at least in principle, obtainable. You do not need to know unknown nuisance parameters or anything. But are the resulting tests valid? What does valid mean?

The formal setup is as in the previous section. We have the data $\bar{X}$ with space $\mathcal{X}$ and set of null distributions $\mathcal{F}_0$. (It may or may not be parameterized.) We also have a finite group $\mathcal{G}$ that acts on the data such that if $\bar{X} \sim F$ for $F \in \mathcal{F}_0$,

$$h_g(\bar{X}) \overset{D}{=} \bar{X} \text{ for any } g \in \mathcal{G}, \quad (20.22)$$
where \( h_g(\mathbf{x}) \) is the action. Let \( G \) be a random element of the group,
\[
G \sim \text{Uniform}(\mathcal{G}) \implies P[G = g] = \frac{1}{\#\mathcal{G}}, \quad g \in \mathcal{G}.
\] (20.23)
(This definition can be extended to \( \mathcal{G} \) being compact.)

**Definition 36.** Let \( T(\mathbf{x}) \) be a given test statistic, and \( \alpha \) a given level. Suppose \( c(\mathbf{x}) \) is a function of \( \mathbf{x} \) that satisfies

1. For each \( \mathbf{x} \in \mathcal{X} \),
\[
P[T(h_G(\mathbf{x})) > c(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}] \leq \alpha, \tag{20.24}
\]
and

2. The function \( c \) is \( G \)-invariant, i.e.,
\[
c(h_g(\mathbf{x})) = c(\mathbf{x}) \quad \text{for all} \quad g \in \mathcal{G}, \ \mathbf{x} \in \mathcal{X}. \tag{20.25}
\]

Then the randomization test rejects the null hypothesis if
\[
T(\mathbf{x}) > c(\mathbf{x}). \tag{20.26}
\]

Part 2 of the definition is virtually automatic, because the randomization distribution of \( T(\mathbf{x}) \) is the same as that for \( T(h_G(\mathbf{x})) \), so the cutoff points for data \( \mathbf{x} \) and \( h_g(\mathbf{x}) \) can be taken to be the same. For example, for the “\( \pm 1 \)” group action in (20.12), the data \( \mathbf{x} = (24.3, 10.9, -4.6, 14.6, 10.0) \) in (20.17) yields the same randomization distribution in (20.18) as does the data \( (-24.3, 10.9, -4.6, 14.6, -10.0) \), or the data \( (24.3, 10.9, 4.6, 14.6, 10.0) \), or the data \( (-24.3, -10.9, -4.6, -14.6, -10.0) \), or any other data set that has the same absolute values, but possibly different signs.

The next result shows that the randomization test really does have the correct level.

**Lemma 35.** Given the above setup, for the test in Definition 36,
\[
P_F[T(\mathbf{X}) > c(\mathbf{x})] \leq \alpha \quad \text{if} \quad \mathbf{X} \sim F \text{ for } F \in \mathcal{F}_0. \tag{20.27}
\]

**Proof.** Look at the random vector
\[
h_G(\mathbf{X}), \tag{20.28}
\]
where \( \mathbf{X} \) and \( G \) are independent, and \( \mathbf{X} \sim F \text{ for } F \in \mathcal{F} \). Then
\[
(h_G(\mathbf{X}) \mid G = g)^D = h_g(\mathbf{X}) =^D \mathbf{X}, \tag{20.29}
\]
where the last equation is the condition (20.22). That is, the conditional distribution of \( h_G(\mathbf{X}) \) given \( G = g \) is the same for each \( g \), hence the unconditional distribution is also the same:
\[
h_G(\mathbf{X}) =^D \mathbf{X}. \tag{20.30}
\]

Thus,
\[
P_F[T(\mathbf{X}) > c(\mathbf{x})] = P_F[T(h_G(\mathbf{X})) > c(h_G(\mathbf{X}))]
= P_F[T(h_G(\mathbf{X})) > c(\mathbf{X})] \quad \text{(by part 2 of Definition 36)}
= E_F[p(\mathbf{X})], \tag{20.31}
\]

where
\[
p(\mathbf{x}) = P[T(h_G(\mathbf{x})) > c(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}]. \tag{20.32}
\]

But by Part 1 of the definition, (20.24), \( p(\mathbf{x}) \leq \alpha \) for all \( \mathbf{x} \). Thus by (20.31), (20.27) holds.

We end this section with some remarks.
20.2. Randomization test formalities

20.2.1 Using p-values

Often in practice, rather than trying to find \( c(\bar{x}) \), it is easier to find the randomization p-value, which for given data \( \bar{x} \) is

\[
p-value(\bar{x}) = P[T(h_G(\bar{x})) \geq T(\bar{x}) \mid X = \bar{x}],
\]

as in (20.21). The p-value is often called the “observed significance level,” because it is basically the level you would obtain if you used the observed test statistic as the cutoff point. The randomization test then rejects the null hypothesis if

\[
p-value(\bar{x}) \leq \alpha.
\]

This procedure works because if \( p-value(\bar{x}) \leq \alpha \), then for the level \( \alpha \) test, it must be that \( c(\bar{x}) < T(\bar{x}) \).

20.2.2 Using the normal approximation

It may be that the Central Limit Theorem can be used to approximate the randomization distribution of the test statistic. For example, consider the statistic \( T(\bar{x}) = \sum x_i \), and the group \( G \) and action as in (20.12). Then

\[
T(h_G(\bar{x})) = \sum \epsilon_i x_i.
\]

The \( \epsilon_i \)'s are iid, with \( P[\epsilon_i = -1] = P[\epsilon_i = +1] = \frac{1}{2} \), hence it is easy to see that as in (20.7),

\[
E[\epsilon_i] = 0 \quad \text{and} \quad Var[\epsilon_i] = 1,
\]

hence

\[
E[T(h_G(\bar{x}))] = \sum E[\epsilon_i] x_i = 0 \quad \text{and} \quad Var[T(h_G(\bar{x}))] = \sum Var[\epsilon_i] x_i^2 = \sum x_i^2.
\]

At least if \( n \) is fairly large and the \( x_i \)'s are not too strange,

\[
\frac{T(h_G(\bar{x})) - E[T(h_G(\bar{x}))]}{\sqrt{Var[T(h_G(\bar{x}))]}} = \frac{\sum \epsilon_i x_i}{\sqrt{\sum x_i^2}} \approx N(0, 1).
\]

An approximate randomization test then uses the normalized version of the test statistic and the normal cutoff, that is, the approximate test rejects if

\[
\frac{\sum x_i}{\sqrt{\sum x_i^2}} > z_\alpha.
\]

This normalized statistic is close to the \( t \)-statistic:

\[
\frac{\sum x_i}{\sqrt{\sum x_i^2}} = \frac{\sqrt{n} \bar{x}}{\sqrt{\sum x_i^2 / n}} \approx \frac{\sqrt{n} \bar{x}}{\sqrt{\sum (x_i - \bar{x})^2 / (n - 1)}} = t,
\]

at least if \( \bar{x} \approx 0 \). But the justification of the cutoff point in (20.39) is based on the randomness of the \( G \), not the randomness of the \( X_i \)'s. In fact, R. A. Fisher used the randomization justification to bolster the use of the normal-theory tests. (He also invented randomization tests.)
20.2.3 \textbf{When the regular and randomization distributions coincide}

It can happen that the randomization distribution of a statistic does not depend on the data $x$. That is, there is some random variable $T^*$ such that for any $x$,

$$(T(h_G(X)) \mid X = x) \overset{D}{=} T^*. \quad (20.41)$$

The conditional distribution of $T(h_G(X))$ given $X = x$ is independent of $x$, which means that the unconditional distribution is the same as the conditional distribution:

$$T(h_G(X)) \overset{D}{=} T^*. \quad (20.42)$$

But by (20.29), $h_G(X)$ and $X$ have the same distribution (under the null), hence

$$T(X) \overset{D}{=} T^*. \quad (20.43)$$

Thus the randomization distribution and the regular distribution are the same, and the randomization cutoff point is the same as the regular cutoff point.

For example, suppose under the null hypothesis that $X_1, \ldots, X_n$ are iid continuous and symmetric about 0, and consider the sign statistic

$$T(x) = \sum \text{Sign}(x_i). \quad (20.44)$$

With the group $G$ being the $\pm 1$'s,

$$T(h_G(x)) = \sum \text{Sign}(\epsilon_i x_i). \quad (20.45)$$

The $\text{Sign}(\epsilon_i x_i)$'s are iid $\text{Uniform}\{-1,+1\}$, no matter what the $x_i$'s are (as long as none are 0). Thus for any $x$,

$$T(h_G(x)) \overset{D}{=} T^* = 2Y - n, \quad (20.46)$$

where $Y \sim \text{Binomial}(n, 1/2)$. But we also know that $T(X) = \sum \text{Sign}(X_i)$ has the same distribution:

$$T(X) \overset{D}{=} 2Y - n. \quad (20.47)$$

That is, the regular and randomization tests coincide.

20.2.4 \textbf{Example: The signed rank test}

Go back to the example with $X_1, \ldots, X_n$ iid $X_i \sim F \in \mathcal{F}$, where $\mathcal{F}$ is the set of continuous distribution functions that have finite variance and are symmetric about their median $\eta$, as in (20.2). Now consider the one-sided testing problem,

$$H_0 : F \in \mathcal{F} \text{ with } \eta = 0 \text{ versus } H_A : F \in \mathcal{F} \text{ with } \eta > 0. \quad (20.48)$$

One drawback to the sign test is that it takes into account only which side of 0 each observation, not how far above or below. It could be that half the data are around -1 and half around 100. The sign statistic would be about 0, yet the data itself would suggest the median is around 50. One modification is to weight the signs with the ranks of the absolute values. That is, let

$$(r_1, \ldots, r_n) = (\text{rank}(|x_1|), \ldots, \text{rank}(|x_n|)), \quad (20.49)$$
20.2. Randomization test formalities

where the smallest \( |x_i| \) has rank 1, the second smallest has rank 2, ..., and the largest has rank \( n \). (Since \( F \) is continuous, we can ignore the possibility of ties.)

For example, with the data from (20.17),

\[
x_1 = 24.3, \ x_2 = 10.9, \ x_3 = -4.6, \ x_4 = 14.6, \ x_5 = 10.0,
\]

(20.50)

the absolute values are

\[
|x_1| = 24.3, \ |x_2| = 10.9, \ |x_3| = 4.6, \ |x_4| = 14.6, \ |x_5| = 10.0,
\]

(20.51)

and the ranks of the absolute values are

\[
r_1 = 5, \ r_2 = 3, \ r_3 = 1, \ r_4 = 4, \ r_5 = 2.
\]

(20.52)

The signed rank statistic is defined as

\[
U(\bar{x}) = \sum r_i \text{Sign}(x_i).
\]

(20.53)

Using the same \( \pm 1 \) group, the randomization statistic is

\[
U(h_G(\bar{x})) = \sum r_i \text{Sign}(\epsilon_i x_i),
\]

(20.54)

because the \( |\epsilon_i x_i| = |x_i| \), hence the ranks are invariant under the group. Notice that this statistic is just the integers 1, ..., \( n \) multiplied by random \( \pm 1 \)'s:

\[
U(h_G(\bar{x})) = D \sum i \epsilon_i.
\]

(20.55)

In particular, the distribution is independent of \( \bar{x} \), which means that the randomization distribution and the regular null distribution are the same (as in Section 20.2.3.) It is easy enough to write out the distribution if \( n \) is small, as in (20.18), but with 1, 2, 3, 4, 5 for the data. It is also easy to simulate. We will take the normalization route from Section 20.2.2. The \( \epsilon_i \)'s are iid \( \text{Uniform}\{ -1, +1 \} \), hence as in (20.36). they have means 0 and variances 1, and as in (20.37),

\[
E[\sum i \epsilon_i] = 0 \quad \text{and} \quad \text{Var}[\sum i \epsilon_i] = \sum i^2 = \frac{n(n+1)(2n+1)}{6}.
\]

(20.56)

Thus the test using the normalized statistic rejects the null hypothesis when

\[
\frac{\sum r_i \text{Sign}(x_i)}{\sqrt{n(n+1)(2n+1)/6}} > z_{\alpha}.
\]

(20.57)

This test is both the randomization test and the regular test.

For example, look at the tire wear data in (20.20). Applying the signed rank test to the \( Z_i \)'s column, note that only one, the \(-0.5\), is negative. It has the second lowest absolute value, so that \( U \) is the sum of 1, 3, ..., 16 minus 2, which equals 132. The variance in (20.56) is 1496, hence the normalized statistic is

\[
\frac{132}{\sqrt{1496}} = 3.413.
\]

(20.58)

which exceeds 1.645, so we would reject the null hypothesis at level 0.05. It looks like the two measurements do differ.
20.3 Two-sample tests

The Normal-based two-sample testing situation has data \(X_1, \ldots, X_n\) iid \(N(\mu_X, \sigma^2)\) and \(Y_1, \ldots, Y_m\) iid \(N(\mu_Y, \sigma^2)\), where the \(X_i\)'s and \(Y_i\)'s are independent, and one tests

\[
H_0 : \mu_X = \mu_Y, \sigma^2 > 0 \text{ versus } H_A : \mu_X > \mu_Y, \sigma^2 > 0.
\]  

(20.59)

(Or, one could have a two-sided alternative.) There are various nonparametric analogs of this problem. The one we will deal with has

\[
X_1, \ldots, X_n \text{ iid } F_X; \ Y_1, \ldots, Y_m \text{ iid } F_Y,
\]  

(20.60)

where the \(X_i\)'s and \(Y_i\)'s are independent, and \(F_X\) and \(F_Y\) are continuous. The null hypothesis is that the distribution functions are equal, and the alternative is that \(F_X\) is stochastically larger than \(F_Y\):

**Definition 37.** The distribution function \(F\) is **stochastically larger** than the distribution function \(G\), written

\[
F >_{\text{st}} G,
\]

if

\[
F(c) \leq G(c) \quad \text{for all } c \in \mathbb{R}, \text{ and}
\]

\[
F(c) < G(c) \quad \text{for some } c \in \mathbb{R}.
\]

(20.62)

It looks like the inequality is going the wrong way, but the idea is that if \(X \sim F\) and \(Y \sim G\), then \(F\) being stochastically larger than \(G\) means that the \(X\) tends to be larger than the \(Y\), or

\[
P[X > c] > P[Y > c], \text{ which implies that } 1 - F(c) > 1 - G(c) \implies F(c) < G(c).
\]

(20.63)

On can also say that "\(X\) is stochastically larger than \(Y\)."

For example, if \(X\) and \(Y\) are both from the same location family, but \(X\)'s parameter is larger than \(Y\)'s, then \(X\) is stochastically larger than \(Y\). The same is true of both are from a distribution with monotone likelihood ratio.

Back to the testing problem. With the data in (20.60), the hypotheses are

\[
H_0 : F_X = F_Y, \text{ continuous } \text{ versus } H_A : F_X >_{\text{st}} F_Y, \ F_X \text{ and } F_Y \text{ continuous.}
\]

(20.64)

We reject the null if the \(x_i\)'s are too much larger than the \(y_i\)'s in some sense. Some possible test statistics:

**Difference in means:** \(A(x, y) = \overline{x} - \overline{y}\)

**Difference in medians:** \(M(x, y) = \text{Median}\{x_1, \ldots, x_n\} - \text{Median}\{y_1, \ldots, y_m\}\)

Wilcoxon \(W:\) \(W(x, y) = \overline{r}_X - \overline{r}_Y\)

Mann-Whitney \(U:\) \(U(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Sign}(x_i - y_i)\)

(20.65)
For the Wilcoxon statistic, we first find the ranks of the data, considered as one big sample of \( n + m \) observations:

\[
\begin{align*}
  r_1 &= \text{rank}(x_1), r_2 = \text{rank}(x_2), \ldots, r_n = \text{rank}(x_n), \\
  r_{n+1} &= \text{rank}(y_1), r_{n+2} = \text{rank}(y_2), \ldots, r_{n+m} = \text{rank}(y_m).
\end{align*}
\]  

(20.66)

Then \( \bar{r}_X \) is the mean of the ranks of the \( x_i \)'s, etc., i.e.,

\[
\bar{r}_X = \frac{1}{n} \sum_{i=n+1}^{n+m} r_i, \quad \bar{r}_Y = \frac{1}{m} \sum_{i=n+1}^{n+m} r_i.
\]  

(20.67)

The Wilcoxon statistic is the same as the difference in means, except with the actual observations replaced by their ranks. The Mann-Whitney statistic counts +1 if an \( x_i \) is larger than a \( y_j \), and –1 if the reverse. Thus the larger the \( x_i \)'s, the more +1’s in the summation.

In order to find the randomization distributions of these statistics, we need to first decide how to randomize. The \( \pm 1 \) group is not relevant here.

### 20.3.1 Permutations

If the null hypothesis is true, then the combined sample of \( n + m \) observations, \( X_1, \ldots, X_n, Y_1, \ldots, Y_m \), is an iid sample. Thus the distribution of the combined sample does not depend on the order of the observations. E.g., if \( n = 2, m = 3 \), the following are true under the null hypothesis:

\[
(X_1, X_2, Y_1, Y_2, Y_3) =^D (X_2, Y_3, Y_2, X_1, Y_1) =^D (Y_3, Y_2, Y_1, X_2, X_1).
\]  

(20.68)

Note that (20.67) is not true under the alternative. The randomization is thus a reordering of the entire sample, which means the group is \( \mathcal{P}_{n+m} \), the group of \( (n + m) \times (n + m) \) permutation matrices. If we let \( \mathbf{z} \) be the \( (n + m) \times 1 \) vector of \( x_i \)'s and \( y_i \)'s,

\[
\mathbf{z} = (x_1, \ldots, x_n, y_1, \ldots, y_m)',
\]  

(20.69)

then the group action for \( \Gamma \in \mathcal{P}_{n+m} \) is

\[
h_\Gamma(\mathbf{z}) = \Gamma \mathbf{z}.
\]  

(20.70)

Thus for any statistic \( T(\mathbf{z}) \), the randomization distribution is the distribution of

\[
T(h_\Gamma(\mathbf{z})) = T(\Gamma \mathbf{z}), \quad \Gamma \sim \text{Uniform}(\mathcal{P}_{n+m}).
\]  

(20.71)

The group has \( (n + m)! \) elements, so for small \( n + m \) it is not too difficult to list out the distribution. But \( (n + m)! \) gets large quite quickly, so simulations or normal approximations are usually useful. (Actually, because the order within the \( x \) and \( y \) groups are irrelevant, there are really only \( \binom{m+n}{n} \) distinct values of \( T(\Gamma \mathbf{z}) \) to worry about.)

The difference in medians test statistic appears to be best done with simulations. I do not see an obvious normal approximation. The other three statistics are easier to deal with, so we will.
20.3.2 Difference in means

Let

\[ a = \begin{pmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \\ -\frac{1}{m} \\ \vdots \\ -\frac{1}{m} \end{pmatrix} = \left( \begin{array}{c} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ -\frac{1}{m} \\ \frac{1}{m} \end{array} \right). \] (20.72)

Then

\[ A(z) = a'z = a'(x_1, \ldots, x_n, y_1, \ldots, y_m)' = \bar{x} - \bar{y} = A(x, y). \] (20.73)

The randomization distribution of the statistic \( A(x, y) \) is thus the distribution of \( A(\Gamma z) = a'\Gamma z \). (20.74)

If \( n \) and \( m \) are not small, we can use a normal approximation. In order to find the mean and variance of \( A(\Gamma z) \), it would be helpful to first find the distribution of the vector \( \Gamma z \).

**Lemma 36.** Suppose \( \psi \) is a fixed \( N \times 1 \) vector, and \( \Gamma \sim \text{Uniform}(P_N) \). Then

\[ E[\Gamma \psi] = \psi 1_N \quad \text{and} \quad \text{Cov}[\Gamma \psi] = s^2_\psi H_N, \] (20.75)

where

\[ s^2_\psi = \frac{\sum (v_i - \overline{v})^2}{N - 1} \quad \text{and} \quad H_N = I_N - \frac{1}{N} 1_N 1_N', \] (20.76)

the \( N \times N \) centering matrix.

The proof is in the next section, Section 20.3.3. The lemma shows that

\[ E[\Gamma z] = \bar{z} 1_{n+m}, \quad \text{where} \quad \bar{z} = \frac{n\bar{x} + m\bar{y}}{n + m}, \] (20.77)

is the pooled mean of both \( x_i \)'s and \( y_i \)'s, and

\[ \text{Cov}[\Gamma z] = s^2_z H_{n+m}, \] (20.78)

where \( s^2_z \) is the sample standard deviation of the combined \( x_i \)'s and \( y_i \)'s. (It is not the pooled standard deviation in the usual sense, because that would subtract the respective means from the observations in the two groups.) Then because \( A(\bar{z}) = a'\bar{z} \), we easily have

\[ E[A(\bar{z})] = \bar{z} a' 1_{n+m} = 0, \] (20.79)

because the sum of the elements of \( a \) is \( m \times (1/m) - n \times (1/n) = 0 \). Also,

\[ \text{Var}[A(\bar{z})] = \text{Var}[a'\Gamma z] \]

\[ = a' \text{Cov}[\Gamma z] a \]

\[ = s^2_z (a' a - 0) \quad \text{(again because \( a' 1_{n+m} = 0 \))} \]

\[ = s^2_z (n \times \frac{1}{n^2} + m \times \frac{1}{m^2}) \]

\[ = s^2_z \left( \frac{1}{n} + \frac{1}{m} \right). \] (20.80)
Thus using the normalized randomization statistic, we reject $H_0$ when

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{1}{n} + \frac{1}{m}}} > z_\alpha. \quad (20.81)$$

If the $s_i^2$ were replaced by the usual pooled variance, then this statistic would be exactly the two-sample $t$ statistic.

20.3.3 Proof of Lemma 36

Let $W = \Gamma \mathbf{v}$. By symmetry, the elements of $W$ all have the same mean, is that $E[W] = a \mathbf{1}_N$ for some $a$. Now

$$1_N' E[W] = a 1_N1_N = Na, \quad (20.82)$$

and

$$1_N' E[WW'] = E[1_N' W] = E[1_N' \Gamma \mathbf{v}] = 1_N' v = \sum v_i, \quad (20.83)$$

because $1_N' \Gamma = 1_N$ for any permutation matrix. Thus $Na = \sum v_i$, or $a = \bar{v}$.

Similarly, the $\text{Var}[W_i]'s$ are all equal, and the covariances $\text{Cov}(W_i, W_j)$ are all equal (if $i \neq j$). Thus for some $b$ and $c$,

$$\text{Cov}[W] = b I_N + c 1_N1_N'. \quad (20.84)$$

Take the trace (the sum of the variances) from both sides:

$$\text{trace}(\text{Cov}[W]) = \text{trace}(E[(W - \bar{v} 1_N)(W - \bar{v} 1_N)'])$$

$$= \text{trace}(E[(\Gamma \mathbf{v} - \bar{v} 1_N)(\Gamma \mathbf{v} - \bar{v} 1_N)'])$$

$$= \text{trace}(\Gamma E[(\bar{v} - \bar{v} 1_N)(\bar{v} - \bar{v} 1_N)']\Gamma')$$

$$= \text{trace}((\bar{v} - \bar{v} 1_N)(\bar{v} - \bar{v} 1_N)')$$

$$= \sum (v_i - \bar{v})^2, \quad (20.85)$$

and

$$\text{trace}(\text{Cov}[W]) = \text{trace}(b I_N + c 1_N1_N'N) = N(b + c), \quad (20.86)$$

hence

$$b + c = \frac{\sum (v_i - \bar{v})^2}{N}. \quad (20.87)$$

Next, note that the sum of the $W_i$'s is a constant, $\sum v_i$, hence its variance is 0. That is,

$$\text{Var}[1_N'W] = \text{Var}[1_N'\Gamma \mathbf{v}] = \text{Var}[1_N' \mathbf{v}] = 0, \quad (20.88)$$

but also

$$0 = \text{Var}[1_N'W] = 1_N \text{Cov}[W] 1_N = 1_N'(b I_N + c 1_N1_N')1_N = Nb + N^2 c. \quad (20.89)$$

Thus

$$b + Nc = 0. \quad (20.90)$$

Solving for $b$ and $c$ in (20.87) and (20.90), we have $c = -b/N$, hence

$$b = \frac{\sum (v_i - \bar{v})^2}{N-1} = s_v^2 \quad \text{and} \quad c = -\frac{1}{N} s_v^2. \quad (20.91)$$
Thus from (20.84),
\[
\text{Cov}[W] = s^2 \mathbf{I}_N - \frac{1}{N} s^2 \mathbf{1}_N \mathbf{1}_N' = s^2 (\mathbf{I}_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N') = s^2 \mathbf{H}_N,
\] (20.92)
as in (20.76).

### 20.3.4 Wilcoxon two-sample rank test

The randomization distribution of the Wilcoxon two-sample test statistic, \( W \) in (20.65), works basically the same as the difference-in-means statistic, but the values are ranks rather than the original variables. That is, we can write
\[
W(z) = W(x, y) = r_X - r_Y = \mathbf{a}' \mathbf{r},
\] (20.93)
where \( \mathbf{r} \) is the \((n + m) \times 1\) vector of ranks from (20.66),
\[
\mathbf{r} = \begin{pmatrix} r_1 \\
\vdots \\
r_n \\
r_{n+1} \\
\vdots \\
r_{n+m} \end{pmatrix} = \begin{pmatrix} \text{rank}(x_1) \\
\vdots \\
\text{rank}(x_n) \\
\text{rank}(y_1) \\
\vdots \\
\text{rank}(y_m) \end{pmatrix}.
\] (20.94)

The randomization distribution is that of \( W(\Gamma z) = \mathbf{a}' \Gamma \mathbf{r} \). Let \( N = n + m \) be the total sample size. Then notice that \( \mathbf{r} \) contains the integers 1, \ldots, \( N \) in some order (assuming no ties among the data). Because \( \Gamma \) is a random permutation, the distribution \( \Gamma \mathbf{r} \) does not depend on the order of the \( r_i \)'s, so that
\[
\Gamma \mathbf{r} = D \Gamma \begin{pmatrix} 1 \\
\vdots \\
N \end{pmatrix} \equiv \Gamma \mathbf{1}_N.
\] (20.95)

In particular, as in Section 20.2.3, the randomization distribution \( W(\Gamma z) \) does not depend on \( z \), hence the randomization distribution is the same as the regular null distribution of \( W(Z) \).

For the normalized statistic, from Lemma 36,
\[
E[\Gamma \mathbf{z}] = E[\Gamma \mathbf{1}_N] = \frac{\sum i=1^N i}{N} \mathbf{1}_N = \frac{N+1}{2} \mathbf{1}_N,
\] (20.96)
and
\[
\text{Cov}[\Gamma \mathbf{z}] = \text{Cov}[\Gamma \mathbf{1}_N] = \frac{\sum i=1^N (i - (N+1)/2)^2}{N-1} \mathbf{H}_N = \frac{N(N+1)}{12} \mathbf{H}_N.
\] (20.97)

Then with the \( \mathbf{a} \) from (20.72), as in (20.79) and (20.80),
\[
E[W(\Gamma \mathbf{z})] = \mathbf{a}' \left( \frac{N+1}{2} \mathbf{1}_N \right) = 0 \quad \text{and}
\]
\[
\text{Cov}[W(\Gamma \mathbf{z})] = \mathbf{a}' \left( \frac{N(N+1)}{12} \mathbf{H}_N \right) \mathbf{a} = \frac{N(N+1)}{12} \left( \frac{1}{n} + \frac{1}{m} \right).
\] (20.98)
The normalized statistic is then
\[
\frac{r_X - r_Y}{\sqrt{\frac{N(N+1)}{12} \left( \frac{1}{n} + \frac{1}{m} \right)}}. \tag{20.99}
\]

**Mann-Whitney two-sample statistic**

It is not immediately obvious, but the Wilcoxon \(W\) and Mann-Whitney \(U\) statistics in (20.65) are equivalent. Specifically,
\[
W(z) = \frac{N}{2} \left( \frac{1}{n} + \frac{1}{m} \right) U(z). \tag{20.100}
\]
Thus they lead to the same test, and this test is often called the **Mann-Whitney/Wilcoxon two-sample test**.

### 20.4 Linear regression \(\rightarrow\) Monotone relationships

Consider the linear regression model,
\[
y_i = \alpha + \beta x_i + e_i. \tag{20.101}
\]
Typically one wishes to test whether \(\beta = 0\), which is (under the usual assumptions) equivalent to testing whether the \(X_i\)'s are independent of the \(Y_i\)'s. If the \(e_i\)'s are iid \(N(0, \sigma^2)\), then \(t\)-tests and \(F\) tests are available. A nonparametric one-sided analog of this situation assumes that \((X_1, Y_1), \ldots, (X_n, Y_n)\) are iid pairs of variables with continuous distributions, and the hypotheses are
\[
H_0 : X_i \text{ is independent of } Y_i, i = 1, \ldots, N \text{ versus } H_A : X_i \text{'s and } Y_i \text{'s are positively related.} \tag{20.102}
\]
The alternative is not very specific, but the idea is larger \(x_i\)'s tend to go with larger \(y_i\)'s. The linear model in (20.101) is a special case if \(\beta > 0\). The alternative includes nonlinear, but increasing, relationships, and makes no assumption about the \(e_i\)'s (other than they are iid and continuous).

#### 20.4.1 Spearman’s \(\rho\)

A natural test statistic in the linear regression model, especially with normal errors, is the sample correlation coefficient,
\[
R(x, y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} = \frac{x' H_n y}{\|H_n x\| \|H_n y\|}. \tag{20.103}
\]
(Recall the \(H_n x\) subtracts \(\overline{x}\) from all the elements of \(x\).) A robustified version uses the rank transform idea, replacing the \(x_i\)'s with their ranks, and the \(y_i\)'s with theirs. In this case, the \(x_i\)'s are ranked within themselves, as are the \(y_i\)'s. (Unlike in the two-sample case, where they were combined.) Then we have two vectors of ranks.
\[
\begin{align*}
\mathbf{r}^x &= \begin{pmatrix}
\text{rank}(x_1 \text{ among } x_1, \ldots, x_n) \\
\vdots \\
\text{rank}(x_n \text{ among } x_1, \ldots, x_n)
\end{pmatrix},
\mathbf{r}^y &= \begin{pmatrix}
\text{rank}(y_1 \text{ among } y_1, \ldots, y_n) \\
\vdots \\
\text{rank}(y_n \text{ among } y_1, \ldots, y_n)
\end{pmatrix}.
\end{align*} \tag{20.104}
\]
Spearman’s $\rho$ is then defined to be the correlation coefficient of the ranks:

$$S(x, y) = \frac{r^x \Gamma_r y}{\|H_n r^x\| \|H_n r^y\|}. \quad (20.105)$$

Under the null hypothesis, the $x_i$’s are independent of the $y_i$’s, hence permuting the $y_i$’s among themselves does not change the distribution. Neither does permuting the $x_i$’s among themselves. That is, if $\Gamma_1$ and $\Gamma_2$ are any $n \times n$ permutation matrices, and the null hypothesis is true, then

$$(X, Y) \overset{D}{=} (X, \Gamma_2 Y) \overset{D}{=} (\Gamma_1 X, Y) = \overset{D}{=} (\Gamma_1 X, \Gamma_2 Y). \quad (20.106)$$

Saying it another way, we can match up the $x_i$’s with the $y_i$’s in any arrangement. We will just permute the $y_i$’s, so that the randomization distribution of any statistic $T(x, y)$ is the distribution of $T(x, \Gamma y)$. For Spearman’s $\rho$, we have

$$S(x, \Gamma y) = \frac{r^x \Gamma r y}{\|H_n r^x\| \|H_n \Gamma r^y\|}, \quad (20.107)$$

because permuting the $y_i$’s has the effect of similarly permuting their ranks. Now, assuming no ties, both $r^x$ and $r^y$ are permutations of the integers $1, \ldots, n$, and is $\Gamma y$, hence

$$\|H_n r^x\|^2 = \|H_n \Gamma r^y\|^2 = \|H_n 1:n\| = \sum (i - (n + 1)/2)^2 = \frac{n(n^2 - 1)}{12}, \quad (20.108)$$

the last equation being from (20.97). Thus the denominator of Spearman’s $\rho$ does not depend on the $\Gamma$. Also, $\Gamma y \overset{D}{=} \Gamma 1:n$, hence

$$S(x, \Gamma y) \overset{D}{=} \frac{1:n}n \Gamma \frac{1:n}n = \frac{1:n}n \Gamma \frac{1:n}n \frac{n(n^2 - 1)}{12}. \quad (20.109)$$

Once again, the randomization distribution does not depend on the data, hence the randomization and regular distributions coincide.

To normalize the statistic, we first find the expected value and covariance of $\Gamma 1:n$, which we did in (20.96) and (20.97):

$$E[\Gamma 1:n] = \frac{n + 1}{2} 1_n \quad \text{and} \quad \text{Cov}[\Gamma 1:n] = \frac{n(n + 1)}{12} H_n. \quad (20.110)$$

Then

$$E[H_n \Gamma 1:n] = \frac{n + 1}{2} H_n 1_n = 0_n \quad \text{and} \quad \text{Cov}[H_n \Gamma 1:n] = \frac{n(n + 1)}{12} H_n H_n H_n^\prime = \frac{n(n + 1)}{12} H_n. \quad (20.111)$$

Finally, the numerator of $S$ has

$$E[(1:n)\Gamma H_n \Gamma 1:n] = 0 \quad (20.112)$$

and

$$\text{Var}[(1:n)\Gamma H_n \Gamma 1:n] = \frac{n(n + 1)}{12} (1:n)\Gamma H_n 1:n = \frac{n(n + 1)}{12} \frac{n(n^2 - 1)}{12}, \quad (20.113)$$
20.5. Fisher’s exact test for independence

\[ E[S(x, y)] = 0 \quad \text{and} \quad \text{Var}[S(x, y)] = \frac{n(n+1)}{12} = \frac{1}{n-1}. \quad (20.114) \]

Thus the approximate level \( \alpha \) test rejects the null hypothesis when
\[ \sqrt{n-1} S(x, y) > z_{\alpha}. \quad (20.115) \]

20.4.2 Kendall’s \( \tau \)

An alternative to Spearman’s \( \rho \) is Kendall’s \( \tau \), which is similar to a correlation coefficient in that it goes from \(-1\) to \(+1\), where close to \(+1\) means there is a strong positive relationship between the \( x_i \)'s and \( y_i \)'s, and \(-1\) means a strong negative relationship. If they are independent, Kendall’s \( \tau \) should be about 0. This statistic is based upon the signs of the slopes comparing pairs of points \((x_i, y_i)\) and \((x_j, y_j)\). Specifically, it looks at the quantities
\[ \text{Sign}(y_i - y_j) \text{Sign}(x_i - x_j). \quad (20.116) \]

This value is \(+1\) if the line segment connecting \((x_i, y_i)\) and \((x_j, y_j)\) is positive, and \(-1\) if the slope is negative. (Why?) We look at that quantity for each pair of points, and add them up:
\[ U(x, y) = \sum \sum_{1 \leq i < j \leq n} \text{Sign}(y_i - y_j) \text{Sign}(x_i - x_j), \quad (20.117) \]

which is the number of positive slopes minus the number of negative slopes. Thus if the \( x_i \)'s and \( y_i \)'s are independent, about half should be positive and half negative, yielding a \( U \) around 0. If all slopes are positive, then the \( x_i \)'s and \( y_i \)'s are perfectly positively related, hence \( U(x, y) = \binom{n}{2} \) (there being \( \binom{n}{2} \) pairs). Similarly, \( U(x, y) = -\binom{n}{2} \) is the \( x_i \)'s and \( y_i \)'s are perfectly negatively related. So Kendall’s \( \tau \) is normalized to be between \( \pm 1 \):
\[ K(x, y) = \frac{\sum \sum_{1 \leq i < j \leq n} \text{Sign}(y_i - y_j) \text{Sign}(x_i - x_j)}{n(n-1)/2}. \quad (20.118) \]

Under the null hypothesis of independence, if the distributions of the variables are continuous, it can be shown that
\[ E[U(X, Y)] = 0 \quad \text{and} \quad \text{Var}[U(X, Y)] = \binom{n}{2} + \frac{2}{3} \binom{n}{3}, \quad (20.119) \]

hence
\[ E[K(X, Y)] = 0 \quad \text{and} \quad \text{Var}[K(X, Y)] = \frac{2}{9} \frac{2n+5}{n(n-1)}. \quad (20.120) \]

20.5 Fisher’s exact test for independence

Now we turn to testing independence in a \( 2 \times 2 \) contingency table. That is, we have
\[ (X_{11}, X_{12}, X_{21}, X_{22}) \sim \text{Multinomial}(n, (p_{11}, p_{12}, p_{21}, p_{22})), \quad (20.121) \]
where we think of the \(X_{ij}\)'s arranged in a table:

\[
\begin{array}{ccc}
\text{Row 1} & \text{Column 1} & \text{Column 2} \\
X_{11} & X_{12} \\
\text{Row 2} & X_{21} & X_{22} \\
\text{Sum} & X_1 & X_2 \\
\end{array}
\]

The null hypothesis is that the rows are independent of the columns, that is, if we let

\[
\begin{align*}
    r_1 &= p_{11} + p_{12}, \\
    r_2 &= p_{21} + p_{22}, \\
    c_1 &= p_{11} + p_{21}, \\
    c_2 &= p_{12} + p_{22},
\end{align*}
\]

so that \(r_1 + r_2 = 1\) and \(c_1 + c_2 = 1\), then

\[H_0: p_{ij} = r_i c_j \text{ for all } i, j \text{ versus } H_A: p_{ij} \neq r_i c_j \text{ for some } i, j.\]

This null hypothesis is not simple, so the distribution of any statistic will likely depend on unknown parameters, even under the null hypothesis. Fisher’s idea was to look at the conditional distribution given the marginals, which are the \(X_{ii}\)'s and \(X_{ij}\)'s. If you know the marginals, then by knowing just \(X_{11}\) one can figure out all the other elements. Thus what we want is the conditional distribution of \(X_{11}\) given the marginals:

\[X_{11} | X_1 = b, X_1 = m.\]

(We left out the other two marginals, since we know \(X_2 = n - X_1\) and \(X_2 = n - X_1\).)

First, we need the conditional space of \(X_{11}\). It is easiest to start by writing out the table with these conditionings:

\[
\begin{array}{ccc}
\text{Row 1} & \text{Column 1} & \text{Column 2} \\
x_{11} & b - x_{11} \\
\text{Row 2} & m - x_{11} & n - m - b + x_{11} \\
\text{Sum} & m & n - m \\
\end{array}
\]

Then each cell must be nonnegative (of course), so the constraints on \(x_{11}\) are

\[x_{11} \geq 0, \ b - x_{11} \geq 0, \ m - x_{11} \geq 0, \ n - b - m + x_{11} \geq 0,\]

which translates to

\[\max\{0, m + b - n\} \leq x_{11} \leq \min\{b, m\}.\]

The joint pmf of the \(X_{ij}\)'s under the null hypothesis is

\[
f(x; r_1, r_2, c_1, c_2) = \binom{n}{x} r_1^{x_1} r_2^{x_2} c_1^{x_1} c_2^{x_2}.\]
Notice that the marginals are sufficient, which means the conditional distribution of $X_{11}$ given the marginals does not depend on the parameters. So we can ignore that part of the pmf, and look at just the factorials part. Then

$$P[X_{11} = x_{11} \mid X_1, = b, X_1 = m] = \frac{n \choose x_{11} + b - x_{11} + m - x_{11}, n - b - m + x_{11}}{\sum \choose w, b - w, m - w, n - b - m + w},$$  \hspace{1cm} (20.130)

where the sum in the denominator is over $w$ in the conditional space $\max \{0, m + b - n\} \leq w \leq \min \{b, m\}$. We rewrite the numerator:

$$\frac{n \choose x_{11} + b - x_{11} + m - x_{11}, n - b - m + x_{11}}{n!} = \frac{x_{11}! (b - x_{11})! (m - x_{11})! (n - b - m + x_{11})!}{n! b! (n - b)!} = \frac{b! (n - b)! x_{11}! (b - x_{11})! (m - x_{11})! (n - b - m + x_{11})!}{\sum \choose w, b - w, m - w, n - b - m + w}.$$

A combinatorial identity is

$$\sum \choose w, b - w, m - w, n - b - m + w = \choose n, m.$$

Thus

$$P[X_{11} = x_{11} \mid X_1, = b, X_1 = m] = \frac{\choose x_{11} + b - x_{11} + m - x_{11}, n - b - m + x_{11}}{\choose n, m},$$  \hspace{1cm} (20.134)

which is the Hypergeometric pmf with parameters $n, b, m$.

Finally, Fisher’s exact test uses $X_{11}$ as the test statistic, and the Hypergeometric distribution to find the cutoff point or p-value. For example, for the alternative that the rows and columns are positively associated, which means $X_{11}$ (and $X_{22}$) should be relatively large, the (conditional) p-value is

$$P[\text{Hypergeometric}(n, b, m) \geq x_{11}].$$  \hspace{1cm} (20.135)

**Example.** Consider the example in which 107 STAT 100 students were categorized according to homework and exam scores:

<table>
<thead>
<tr>
<th>HW ↓</th>
<th>Exams →</th>
<th>lo</th>
<th>hi</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>lo</td>
<td>38</td>
<td>16</td>
<td></td>
<td>$b = 54$</td>
</tr>
<tr>
<td>hi</td>
<td>16</td>
<td>37</td>
<td></td>
<td>$n - b = 53$</td>
</tr>
<tr>
<td>Sum</td>
<td>$m = 54$</td>
<td>$n - m = 53$</td>
<td></td>
<td>$n = 107$</td>
</tr>
</tbody>
</table>

(It is just a coincidence that $b = m$.) The range of $X_{11}$ conditional on the marginals is $\{1, \ldots , 54\}$ and the Hypergeometric pmf is

$$f(w; 107, 54, 54) = \choose w, 54 \cdot \choose 53 - w, 54, 107.$$  \hspace{1cm} (20.137)
Then the one-sided p-value (for testing positive association) is then

$$P[\text{Hypergeometric}(107, 54, 54) \geq 38] = \sum_{w=38}^{54} f(w; 107, 54, 54) = 0.00003. \quad (20.138)$$

That is a tiny p-value, so we reject the null hypothesis: people who do well on the homework tend to do well on the exams.