Mathematical Statistics
Statistical Inference Notes

By

Daniel Ruiz

Last Updated: November 2020

Contents

1 Probability **1**
- 1.1 Set Theory . . . 1
- 1.2 Basics of Probability Theory . . . 2
  - 1.2.1 Counting . . . 4
- 1.3 Conditional Probability and Independence . . . 4
- 1.4 Random Variables . . . 6
- 1.5 Distribution Functions . . . 6
- 1.6 Density and Mass Functions . . . 7

2 Transformations and Expectations **9**
- 2.1 Distributions of Functions of a Random Variable . . . 9
- 2.2 Expected Values . . . 10
- 2.3 Moments and Moment Generating Functions . . . 11
- 2.4 Differentiating Under an Integral Sign . . . 13

3 Common Families of Distributions **15**
- 3.1 Discrete Distributions . . . 15
- 3.2 Continuous Distributions . . . 18
- 3.3 Exponential Families . . . 21
- 3.4 Location and Scale Families . . . 23
- 3.5 Inequalities and Identities . . . 24
- 3.6 Miscellanea . . . 25

4 Multiple Random Variables **27**
- 4.1 Joint and Marginal Distributions . . . 27
- 4.2 Conditional Distributions and Independence . . . 29
- 4.3 Bivariate Transformations . . . 31
- 4.4 Hierarchical Models and Mixture Distributions . . . 32
- 4.5 Covariance and Correlation . . . 32
- 4.6 Multivariate Distributions . . . 34
- 4.7 Inequalities . . . 37
  - 4.7.1 Numerical Inequalities . . . 37
  - 4.7.2 Functional Inequalities . . . 38

5 Properties of a Random Sample **39**
- 5.1 Basic Concepts of Random Samples . . . 39
- 5.2 Sums of Random Variables from a Random Sample . . . 40
- 5.3 Sampling from the Normal Distribution . . . 42
  - 5.3.1 Properties of the Sample Mean and Variance . . . 42
  - 5.3.2 The Derived Distributions: Student's t and Snedecor's F . . . 43
- 5.4 Order Statistics . . . 44
- 5.5 Convergence Concepts . . . 46
  - 5.5.1 Convergence in Probability . . . 46

5.5.2 Almost Sure Convergence . . . 46
5.5.3 Convergence in Distribution . . . 47
5.5.4 The Delta Method . . . 48
5.6 Generating a Random Sample . . . 50
5.6.1 The Accept / Reject Algorithm . . . 50
5.7 Miscellanea . . . 51

6 Principles of Data Reduction **53**
6.1 The Sufficiency Principle . . . 53
6.1.1 Sufficient Statistics . . . 54
6.1.2 Minimal Sufficient Statistics . . . 54
6.1.3 Sufficient, Ancillary, and Complete Statistics . . . 55
6.2 The Likelihood Principle . . . 56
6.2.1 The Likelihood Function . . . 56
6.2.2 The Formal Likelihood Principle . . . 56
6.3 The Equivariance Principle . . . 57
6.4 Miscellanea . . . 58

7 Point Estimation **61**
7.1 Methods of Finding Estimators . . . 61
7.1.1 Method of Moments . . . 61
7.1.2 Maximum Likelihood Estimators . . . 62
7.1.3 Bayes Estimators . . . 63
7.1.4 The Expectation-Maximization (EM) Algorithm . . . 63
7.2 Methods of Evaluating Estimators . . . 63
7.2.1 Mean Squared Error . . . 63
7.2.2 Best Unbiased Estimators . . . 64
7.2.3 Sufficiency and Unbiasedness . . . 65
7.2.4 Loss Function Optimality . . . 66
7.3 Miscellanea . . . 67

8 Hypothesis Testing **69**
8.1 Introduction . . . 69
8.2 Methods of Finding Tests . . . 70
8.2.1 Likelihood Ratio Tests . . . 70
8.2.2 Bayesian Tests . . . 70
8.2.3 Union-Intersection and Intersection-Union Tests . . . 70
8.3 Methods of Evaluating Tests . . . 71
8.3.1 Error Probabilities and the Power Function . . . 71
8.3.2 Most Powerful Tests . . . 72
8.3.3 Sizes of Union-Intersection and Intersection-Union Tests . . . 74
8.3.4 p-Values . . . 74
8.3.5 Loss Function Optimality . . . 75

Chapter 1

Probability

1.1 Set Theory

Definition 1.1: Sample Space

The sample space Ω *is the set of possible outcomes of an experiment. Points* ω *in* Ω *are called* ***sample outcomes, realizations****, or* ***elements****. Subsets of* Ω *are called* ***events****.*

Definition 1.2: Event

An event is any collection of possible outcomes of an experiment, that is, any subset of S *(including* S *itself).*

Theorem 1.1

For any three events, A, B*, and* C*, defined on a sample space* S*,*

a. Commutativity

$$A \cup B = B \cup A, \tag{1.1}$$
$$A \cap B = B \cap A; \tag{1.2}$$

b. Associativity

$$A \cup (B \cup C) = (A \cup B) \cup C, \tag{1.3}$$
$$A \cap (B \cap C) = (A \cap B) \cap C; \tag{1.4}$$

c. Distributive Laws

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \tag{1.5}$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C); \tag{1.6}$$

d. DeMorgan's Laws

$$(A \cup B)^c = A^c \cap B^c, \tag{1.7}$$
$$(A \cap B)^c = A^c \cup B^c \tag{1.8}$$

Definition 1.3: Mutually Exclusive / Disjoint

*Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \emptyset$. We say that $A_1, A_2, ..$ are **pairwise disjoint** (or **mutually exclusive**) if $A_i \cap A_j = \emptyset$ for all $i \neq j$.*

Definition 1.4: Partition

*A **partition** of S is a sequence of pairwise disjoint sets $A_1, A_2, ...$ such that $\cup_{i=1}^{\infty} A_i = S$.*

1.2 Basics of Probability Theory

Definition 1.5: Sigma Algebra

A collection of subsets of S is called a sigma algebra (or Borel field), denoted by $\mathcal{F}$, if it satisfies the following three properties:

1. $\emptyset \in \mathcal{F}$
2. *If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ ($\mathcal{F}$ is closed under complementation)*
3. *If a sequence of sets $A_1, A_2, ... \in \mathcal{F}$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$. ($\mathcal{F}$ is closed under countable unions)*

Corollary 1.1: Closure Under Intersections

Let $\mathcal{B}$ be a sigma algebra. If $A_1, A_2, ... \in \mathcal{F}$, then $\cap_{i=1}^{\infty} A_i \in \mathcal{B}$. (Closed under countable intersections)

Definition 1.6: Probability Function

Let Ω be a sample space with associated sigma algebra $\mathcal{F}$. A probability function is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ that satisfies

1. $\mathbb{P}(A) \geq 0$ *for every* $A \in \mathcal{F}$.
2. $\mathbb{P}(\Omega) = 1$.
3. *If $A_1, A_2, ...$ are disjoint then*

$$\mathbb{P}\big(\cup_{i=1}^{\infty} A_i\big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \tag{1.9}$$

Theorem 1.2

Let $S = \{s_1, ..., s_n\}$ be a finite set. Let $\mathcal{F}$ be any sigma algebra of subsets of S. Let $p_1, ..., p_n$ be nonnegative numbers that sum to 1. For any $A \in \mathcal{F}$, define $P(A)$ by

$$P(A) = \sum_{i:s_i \in A} p_i. \tag{1.10}$$

(The sum over an empty set is defined to be 0). Then P is a probability function on $\mathcal{F}$. This remains true of $S = \{s_1, s_2, ...\}$ is a countable set.

Theorem 1.3

If P is a probability function and A is any set in $\mathcal{F}$, then

- **a.** $P(\emptyset) = 0$;
- **b.** $P(A) \leq 1$;
- **c.** $P(A^c) = 1 - P(A)$.

Theorem 1.4: Probability Function Properties

If P is a probability function and A and B are any sets in $\mathcal{F}$, then

- **a.** $P(B \cap A^c) = P(B) - P(A \cap B)$
- **b.** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
- **c.** *If $A \subset B$, then $P(A) \leq P(B)$.*

Corollary 1.2: Bonferroni's Inequality

Let A and B be events with P a probability function. Then,

$$P(A \cap B) \geq P(A) + P(B) - 1. \tag{1.11}$$

This inequality only becomes useful if $P(A) + P(B) > 0$ as it can then establish a non-trivial lower bound on $P(A \cap B)$.

Proposition 1.1

If P is a probability function, then

- **a.** $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ *for any partition $C_1, C_2, ...$;*
- **b.** $P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ *for any sets $A_1, A_2, ...$ (Boole's Inequality).*

Definition 1.7: Monotone Increasing / Decreasing

*A sequence of sets $A_1, A_2, ...$ is **monotone increasing** if $A_1 \subset A_2 \subset ...$ and we define $\lim_{n\to\infty} A_n := \cup_{i=1}^{\infty} A_i$. A sequence of sets $A_1, A_2, ...$ is **monotone decreasing** if $A_1 \supset A_2 \supset ...$ and we similarly define $\lim_{n\to\infty} A_n := \cap_{i=1}^{\infty} A_i$. In either case, we write $A_n \to A$.*

Corollary 1.3: Generalized Bonferroni's Inequality

Let $A_1, A_2, ..., A_n \in \mathcal{F}$. Then, we have the inequality

$$P\big(\cap_{i=1}^{n} A_i\big) \geq \sum_{i=1}^{n} P(A_i) - (n-1) \tag{1.12}$$

1.2.1 Counting

Theorem 1.5: Fundamental Theorem of Counting

If a job consists of k separate tasks, the i^{th} of which can be done in n_i ways, $i = 1, ..., k$, then the entire job can be done in $n_1 \times n_2 \times ... \times n_k$ ways.

Proposition 1.2: Counting Methods

The number of possible arrangements of size r from n objects for

1. ***Ordered, Without Replacement*** *is* $\frac{n!}{(n-r)!}$.
2. ***Ordered, With Replacement*** *is* n^r.
3. ***Unordered, Without Replacement*** *is* $\binom{n}{r}$.
4. ***Unordered, With Replacement*** *is* $\binom{n+r-1}{r}$.

Definition 1.8: Uniform Probability Distribution

If Ω is finite and if each outcome is equally likely, then

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}, \tag{1.13}$$

which is called the ***uniform probability distribution****.*

Proposition 1.3

Suppose that we have some ordered arrangement of k values and we are interested in how many possible arrangement there are. If we have m distinct numbers repeated $k_1, k_2, .., k_m$ times for each number, then the number of ordered samples is

$$\frac{k!}{k_1!k_2!...k_m!} \tag{1.14}$$

1.3 Conditional Probability and Independence

Definition 1.9: Conditional Probability

If A, B are events and $\mathbb{P}(B) > 0$, then the ***conditional probability*** *of A given B is*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} \tag{1.15}$$

Lemma 1.1: Properties of Conditional Probability

Let $\mathbb{P}$ be a probability measure over some σ-algebra $\mathcal{F}$. All events considered in the following are members of $\mathcal{F}$.

1. *For any pairs of events A and B,*

$$\mathbb{P}(AB) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) \tag{1.16}$$

2. *$\mathbb{P}(\cdot|B)$ satisfies the axioms of probability measure, for fixed B. However, in general $\mathbb{P}(A|\cdot)$ does not satisfy the axioms of probability, for fixed A.*

3. *In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.*

Theorem 1.6: Law of Total Probability

Let $A_1, ..., A_k$ be a partition of Ω. Then, for any event B, we have that

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i). \tag{1.17}$$

Theorem 1.7: Bayes Theorem

Let $A_1, A_2, ..., A_k$ be a partition of Ω such that $P(A_i) > 0$ for each i. If $P(B) > 0$, then for each $i = 1, ..., k$, we have

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)} \tag{1.18}$$

*We call $\mathbb{P}(A_i)$ the **prior probability of** A and $\mathbb{P}(A_i|B)$ the **posterior probability of** A.*

Definition 1.10: Independent Events

*Two events A and B are **statistically independent** if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{1.19}$$

As a consequence of this, one has that $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

Theorem 1.8

If A and B are independent events, then the following pairs are also independent:

a. *A and B^c,*

b. *A^c and B,*

c. *A^c and B^c*

Definition 1.11: Mutually Independent

A collection of events $A_1, ..., A_n$ are mutually independent if for any subcollection $A_{i_1}, ..., A_{i_k}$, we have

$$\mathbb{P}\big(\cap_{j=1}^{k} A_j\big) = \Pi_{j=1}^{k}\mathbb{P}(A_{i_j}) \tag{1.20}$$

1.4 Random Variables

Definition 1.12: Measurable Map

Let $\mathcal{F}$ be a σ-field. A map $X : \Omega \to \mathbb{R}$ is said to be measurable, if for every $x \in \mathbb{R}$, we have $\{\omega : X(\omega) \leq x\} \in \mathcal{F}$.

Definition 1.13: Random Variable

Let Ω be a sample space. A ***random variable*** *is a measurable mapping*

$$X : \Omega \to \mathbb{R} \tag{1.21}$$

that assigns a real number $X(\omega)$ to each outcome ω.

Definition 1.14: Induced Probability Function

Suppose that we have a finite sample space $S = \{s_1, ..., s_n\}$ with probability function P. Let X be a random variable with range $\chi = \{x_1, ..., x_m\}$. We define a probability function P_X on χ in the following way:

$$P_X(X = x_i) := P\big(\{s_j \in S : X(s_j) = x_i\}\big). \tag{1.22}$$

The function P_X is referred to as the induced probability function on χ.

Definition 1.15: Induced Probability Function (Uncountable Case)

Extending the definition of induced probability function to the case where χ is countable is straightforward. Let S be a sample space. If χ is uncountable, then we define the induced probability function P_X for any set $A \subset \chi$ as follows:

$$P_X(X \in A) := P\big(\{s \in S : X(s) \in A\}\big) \tag{1.23}$$

1.5 Distribution Functions

Definition 1.16: Cumulative Distribution Function (CDF)

The cumulative distribution function or CDF of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P_X(X \leq x), \text{ for all } x \tag{1.24}$$

Theorem 1.9

The function $F(x)$ is a CDF if and only if the following three conditions hold:

a. $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$.

b. $F(x)$ *is nondecreasing function of* x.

c. $F(x)$ *is right-continuous; that is, for every number* x_0, $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

Definition 1.17: Continuous and Discrete Random Variables

A random variable X is continuous if $F_X(x)$ is a continuous function of x. A random variable X is discrete if $F_X(x)$ is a step function of x.

Definition 1.18: Identically Distributed

The random variables X and Y are identically distributed if, for every set $A \in \mathcal{F}$, $P(X \in A) = P(Y \in A)$.[a]

[a] We note that this definition does not say that $X = Y$

Theorem 1.10: CDF $\leftrightarrow$ Identical Distribution

Let X and Y be two random variables. The following two statements are equivalent:

- **a.** *The random variables X and Y are identically distributed.*
- **b.** *$F_X(x) = F_Y(y)$ for every x.*

1.6 Density and Mass Functions

Definition 1.19: Probability Mass Function (PMF)

The probability mass function (PMf) of a discrete random variable X is given by

$$f_X(x) = P(X = x) \text{ for all } x \tag{1.25}$$

Definition 1.20: Probability Density Function (PDF)

The probability density function (PDF), $f_X(x)$ of a continuous random variable X is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \ \forall x \in \mathbb{R} \tag{1.26}$$

We note that the relationship (1.26) does not always hold as $F_X(x)$ may be continuous but not differentiable. There exist continuous random variables for which the integral relationship does not exist for any $f_X(x)$. In this text, we assume that (1.26) holds for any continuous random variable.

Definition 1.21: Absolutely Continuous Random Variable

In more advanced texts, a random variable is called absolutely continuous if (1.26) holds.

Notation 1.1: $X \sim Y$

The expression "X has a distribution given by $F_X(x)$" is abbreviated symbolically by "$X \sim F_X(x)$", where we read the symbol "$\sim$" as "is distributed as". We can similarly write $X \sim f_X(x)$ or, if X and Y have the same distribution, $X \sim Y$.

Theorem 1.11

A function $f_X(x)$ is a PDF (or PMF) of a random variable X if and only if

a. $f_X(x) \geq 0$ *for all* x.

b. $\sum_x f_X(x) = 1$ *(PMF) or* $\int_{-\infty}^{\infty} f_X(x)dx = 1$ *(PDF).*

Chapter 2

Transformations and Expectations

2.1 Distributions of Functions of a Random Variable

Transforming a Random Variable

Let X be a random variable with CDF $F_X(x)$, then any function of X, say $g(X)$ is also a random variable. Often $g(X)$ is of interest itself and we write $Y = g(X)$ to denote the new random variable $g(X)$. We can describe probabilistic behaviour of Y in terms of X:

$$P(Y \in A) = P(g(X) \in A). \tag{2.1}$$

Let $\mathcal{X}$ be the range of X and $\mathcal{Y}$ be the range of Y[a]. The function acts as

$$g : \mathcal{X} \to \mathcal{Y} \tag{2.2}$$

We associate with g an inverse mapping, denoted by $g^{-1} : \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{F})$ denotes the power set of $\mathcal{F}$. It is defined by

$$g^{-1}(A) = \{x \in \mathcal{X} : g(x) \in A\} \text{ for } A \in \mathcal{P}(\mathcal{X}) \tag{2.3}$$

If A is a point set such as $A = \{y\}$, then

$$g^{-1}(\{y\}) = \{x \in \mathcal{X} : g(x) = y\}. \tag{2.4}$$

In this case, we often write $g^{-1}(y)$ instead of $g^{-1}(\{y\})$. With $Y = g(X)$, we can write for any set $A \subset \mathcal{Y}$,

$$P(Y \in A) = P(X \in g^{-1}(A)) \tag{2.5}$$

[a] The author calls $\mathcal{X}$ and $\mathcal{Y}$ the sample space of X and Y respectively, which seems inconsistent with introduced terminology.

Definition 2.1: Support Set

When the transformation is from X to $Y = g(X)$, it is most convenient to use

$$\mathcal{X} = \{x : f_X(x) > 0\} \qquad \text{and} \qquad \mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}. \tag{2.6}$$

The PDF of the random variable X is positive only on the set $\mathcal{X}$ and is 0 elsewhere. Such a set is called the support set of a distribution.

Theorem 2.1

Let X have CDF $F_X(x)$, let $Y = g(X)$, and let $\mathcal{X}$ and $\mathcal{Y}$ be defined as in (2.6).

- **a.** *If g is an increasing function on $\mathcal{X}$, then $F_Y(y) = F_X\big(g^{-1}(y)\big)$ for $y \in \mathcal{Y}$.*
- **b.** *If g is a decreasing function on $\mathcal{X}$ and X is a continuous random variable, then $F_Y(y) = 1 - F_X\big(g^{-1}(y)\big)$ for $y \in \mathcal{Y}$.*

Theorem 2.2

Let X have PDF $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Let $\mathcal{X}$ and $\mathcal{Y}$ be defined by (2.6). Suppose that $f_X(x)$ is continuous on $\mathcal{X}$ and that $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then, the PDF of Y is given by

$$f_Y(y) = \begin{cases} f_X\big(g^{-1}(y)\big)\left|\frac{d}{dy}g^{-1}(y)\right| & y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

Theorem 2.3

Let X have PDF $f_X(x)$, let $Y = g(X)$, and define the "sample space" $\mathcal{X}$ as in (2.6). Suppose there exists a partition $A_0, A_1, ..., A_k$ of $\mathcal{X}$ such that $P(X \in A_0) = 0$ and $f_X(x)$ is continuous on each A_i. Further, suppose there exist functions $g_1(x), ..., g_k(x)$ defined on $A_1, ..., A_k$, respectively, satisfying

- ***i.*** *$g(x) = g_i(x)$, for $x \in A_i$,*
- ***ii.*** *$g_i(x)$ is monotone on A_i,*
- ***iii.*** *The set $\mathcal{Y} = \{y : y = g_i(x) \text{ for some } x \in A_i\}$ is the same for each $i = 1, ..., k$, and*
- ***iv.*** *$g_i^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$, for each $i = 1, ..., k$.*

Then

$$f_Y(y) = \begin{cases} \sum_{i=1}^{k} f_X\big(g_i^{-1}(y)\big)\left|\frac{d}{dy}g_i^{-1}(y)\right| & y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

Theorem 2.4: Probability Integral Transformation

Let X have continuous CDF $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0, 1)$, that is, $P(Y \leq y) = y, 0 < y < 1$.

2.2 Expected Values

Definition 2.2: Expected Value / Mean

Let X be a random variable. The expected value, mean or expectation of a random variable $g(X)$, denoted $E[g(X)]$, is

$$E[g(X)] = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x) f_X(x) = \sum_{x \in \mathcal{X}} g(x) P(X = x) & \text{if } X \text{ is discrete} \end{cases} \tag{2.9}$$

Theorem 2.5: Expectation Properties

Let X be a random variable and let a, b and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,

- **a.** $E[a \cdot g_1(X) + b \cdot g_2(X) + c] = aE[g_1(X)] + bE[g_2(X)] + c$.
- **b.** *If $g_1(x) \geq 0$ for all x, then $E[g_1(X)] \geq 0$.*
- **c.** *If $g_1(x) \geq g_2(x)$ for all x, then $E[g_1(X)] \geq E[g_2(X)]$.*
- **d.** *If $a \leq g_1(x) \leq b$ for all x, then $a \leq E[g_1(X)] \leq b$.*

2.3 Moments and Moment Generating Functions

Definition 2.3: Moments and Central Moments

For each integer n, the n^{th} moment of X (or $F_X(x)$), μ'_n is

$$\mu'_n = E[X^n]. \tag{2.10}$$

The n^{th} central moment of X, μ_n is

$$\mu_n = E[(X - \mu)^n], \tag{2.11}$$

where $\mu = \mu'_1 = E[X]$.

Definition 2.4: Variance & Standard Deviation

The variance of a random variable X is its second central moment, $Var[X] = E[(X - E[X])^2]$. The positive square root of $Var[X]$ is the standard deviation of X.

Theorem 2.6

If X is a random variable with finite variance, then for any constants a and b,

$$Var[aX + b] = a^2 Var[X] \tag{2.12}$$

Definition 2.5: Moment Generating Function (MGF)

Let X be a random variable with CDF F_X. The moment generating function (MGF) of X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) = E[e^{tX}] \tag{2.13}$$

provided that the expectation exists for t in some neighbourhood of 0. That is, there exists an $h > 0$ such that, for all $-h < t < h$, $E[e^{tX}]$ exists. If such a neighbourhood does not exist, we say that the moment generating function does not exist.

Proposition 2.1

Let X be a random variable. We can write the MGF of X as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx \qquad \text{if } X \text{ is continuous} \tag{2.14}$$

or

$$M_X(t) = \sum_x e^{tx} P(X = x) \qquad \text{if } X \text{ is discrete} \tag{2.15}$$

Theorem 2.7

If X has MGF $M_X(t)$, then

$$E[X^n] = M_X^{(n)}(0), \tag{2.16}$$

where we define

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)\big|_{t=0}. \tag{2.17}$$

That is, the n^{th} moment is equal to the n^{th} derivative of $M_X(t)$ evaluated at $t = 0$.

Theorem 2.8

Let $F_X(x)$ and $F_Y(y)$ be two CDFs all of whose moments exist.

- **a.** *If X and Y have bounded support, then $F_X(u) = F_Y(u)$ for all u if and only if $E[X^r] = E[Y^r]$ for all integers $r = 0, 1, 2, ...$*
- **b.** *If the moment generating functions exist and $M_X(t) = M_Y(t)$ for all t in some neighbourhood of 0, then $F_X(u) = F_Y(u)$ for all u.*

Theorem 2.9: Convergence of MGFs

Suppose that $\{X_i; i = 1, 2, ...\}$ is a sequence of random variables, each with MGF $M_{X_i}(t)$. Furthermore, suppose that

$$\lim_{i\to\infty} M_{X_i}(t) = M_X(t), \qquad \text{for all } t \text{ in a neighbourhood of } 0, \tag{2.18}$$

and $M_X(t)$ is an MGF. Then there is a unique CDF F_X whose moments are determined by $M_X(t)$ and, for all x where $F_X(x)$ is continuous, we have

$$\lim_{i\to\infty} F_{X_i}(x) = F_X(x). \tag{2.19}$$

That is, convergence, for $|t| < h$, of MGFs to an MGF implies convergence of CDFs.

Lemma 2.1

Let $a_1, a_2, ...$ be a sequence of numbers converging to a, that is, $\lim_{n\to\infty} a_n = a$. Then

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a \tag{2.20}$$

Theorem 2.10

For any constants a and b the MGF of the random variable $aX + b$ is given by

$$M_{aX+b}(t) = e^{bt} M_X(at) \tag{2.21}$$

2.4 Differentiating Under an Integral Sign

Theorem 2.11: Leibnitz's Rule

If $f(x, \theta)$, $a(\theta)$, and $b(\theta)$ are differentiable with respect to θ, then

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)dx = f\big(b(\theta),\theta\big)\frac{d}{d\theta}b(\theta) - f\big(a(\theta),\theta\big)\frac{d}{d\theta}a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta} f(x,\theta)dx \tag{2.22}$$

In the case where $a(\theta)$ and $b(\theta)$ are constant, we have a special case of Leibnitz's Rule:

$$\frac{d}{d\theta}\int_{a}^{b} f(x,\theta)\mathrm{d}x = \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta} f(x,\theta)\mathrm{d}x \tag{2.23}$$

However, if the range of integration is infinite, problems can arise. To ensure that we can interchange integration and differentiation order in these cases, we explore the conditions under which one can safely do this.

Theorem 2.12

Suppose the function $h(x, y)$ is continuous at y_0 for each x, and there exists a function $g(x)$ satisfying

- ***i.*** *$|h(x, y)| \leq g(x)$ for all x and y,*
- ***ii.*** *$\int_{-\infty}^{\infty} g(x)dx < \infty$.*

Then

$$\lim_{y\to y_0}\int_{-\infty}^{\infty} h(x,y)dx = \int_{-\infty}^{\infty}\lim_{y\to y_0} h(x,y)dx \tag{2.24}$$

Theorem 2.13: Validity of Differentiating under Integral

Suppose $f(x, \theta)$ is differentiable at $\theta = \theta_0$, that is,

$$\lim_{\delta\to 0}\frac{f(x,\theta_0+\delta) - f(x,\theta_0)}{\delta} = \frac{\partial}{\partial\theta} f(x,\theta)\big|_{\theta=\theta_0} \tag{2.25}$$

exists for every x, and there exists a function $g(x, \theta_0)$ and a constant $\delta_0 > 0$ such that

- ***i.*** *$\left|\frac{f(x,\theta_0+\delta)-f(x,\theta_0)}{\delta}\right| \leq g(x,\theta_0)$, for all x and $|\delta| \leq \delta_0$,*
- ***ii.*** *$\int_{-\infty}^{\infty} g(x,\theta_0)dx < \infty$.*

Then

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)dx\big|_{\theta=\theta_0} = \int_{-\infty}^{\infty}\left[\frac{\partial}{\partial\theta} f(x,\theta)\big|_{\theta=\theta_0}\right]dx \tag{2.26}$$

It's important to realize the statement of the above theorem is for one value of θ. It is for each value θ_0 for which $f(x,\theta)$ is differentiable at θ_0 and satisfies conditions (i) and (ii), can the order of integration and differentiation be interchanged. In the following corollary, we provide a condition on interchanging integration and differentiation over a differentiable range of θ.

Corollary 2.1

Suppose that $f(x,\theta)$ is differentiable in θ and there exists a function $g(x,\theta)$ satisfying

$$\left|\frac{\partial}{\partial\theta}f(x,\theta)\Big|_{\theta=\theta'}\right| \leq g(x,\theta) \qquad \text{for all } \theta' \text{ such that } |\theta'-\theta|\leq\delta_0 \tag{2.27}$$

and $\int_{-\infty}^{\infty} g(x,\theta)dx < \infty$. Then,

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)dx = \int_{-\infty}^{\infty}\left[\frac{\partial}{\partial\theta}f(x,\theta)\right]dx \tag{2.28}$$

Definition 2.6: Uniform Convergence

Suppose that E is a set and $(f_n)_{n\in\mathbb{N}}$ is a sequence of real-valued functions on it. We say that the sequence $(f_n)_{n\in\mathbb{N}}$ is uniformly convergent on E with limit $f: E\to\mathbb{R}$ if for every $\epsilon>0$, there exists a natural number N such that for all $n\geq N$ and $x\in E$

$$|f_n(x)-f(x)|<\epsilon. \tag{2.29}$$

Definition 2.7: Convergence of Series

Let $g=\sum_{n=1}^{\infty} f_n$ and $s_n(x)=\sum_{j=1}^{n} f_j(x)$ be the partial sums. We say that g converges

i. ***pointwise*** *on E if and only if the sequence of partial sums s_n converges for every $x\in E$.*

ii. ***uniformly*** *on E if and only if s_n converges uniformly as $n\to\infty$.*

iii. ***absolutely*** *on E if and only if $\sum_{n=1}^{\infty}|f_n|$ converges for every $x\in E$.*

Theorem 2.14

Suppose that the series $\sum_{x=0}^{\infty} h(\theta,x)$ converges for all θ in an interval (a,b) of real numbers and

i. *$\frac{\partial}{\partial\theta}h(\theta,x)$ is continuous in θ for each x,*

ii. *$\sum_{x=0}^{\infty}\frac{\partial}{\partial\theta}h(\theta,x)$ converges uniformly on every closed bounded subinterval of (a,b).*

Then

$$\frac{d}{d\theta}\sum_{x=0}^{\infty} h(\theta,x) = \sum_{x=0}^{\infty}\frac{\partial}{\partial\theta}h(\theta,x) \tag{2.30}$$

Theorem 2.15

Suppose the series $\sum_{x=0}^{\infty} h(\theta,x)$ converges uniformly on $[a,b]$ and that, for each x, $h(\theta,x)$ is a continuous function of θ. Then

$$\int_a^b\sum_{x=0}^{\infty} h(\theta,x)d\theta = \sum_{x=0}^{\infty}\int_a^b h(\theta,x)d\theta \tag{2.31}$$

Chapter 3

Common Families of Distributions

3.1 Discrete Distributions

Notation 3.1: Parametric Distribution Parameters $P(\mathbf{X}|\theta)$

When we are dealing with parametric distributions, as will almost always be the case, the distribution is dependent on values of the parameters. In order to emphasize this fact, we write them in the PMF *preceded by | symbol (given). This convention is also used with* CDFS*,* PDFS*, expectations, and other places where it might be necessary to keep track of the parameters.*

Definition 3.1: Uniform Distribution (Discrete)

A random variable X has a discrete uniform $(1, N)$ distribution if

$$P(X = x|N) = \frac{1}{N}, \qquad x = 1, 2, ..., N, \tag{3.1}$$

where N is a specified integer.

Proposition 3.1: Properties of Uniformly Distributed Variables

Let X be uniformly distributed on $(1, N)$. Then,

1. $E[X] = \frac{N+1}{2}$.
2. $Var[X] = \frac{(N+1)(N-1)}{12}$.

Definition 3.2: Hypergeometric Distribution

The hypergeometric distribution has many applications in finite population sampling. Suppose that we have a collection of N objects, comprised of two populations, labeled A and B. Let M be the size of the A population. If we let X denote the number of A objects chosen in a sample of size K, then X has a hypergeometric distribution given by

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, \qquad x = 0, 1, ..., K \tag{3.2}$$

Proposition 3.2: Properties of Hypergeometric Distributed Variables

Let $X \sim hypergeometric(N, M, K)$, then

1. $E[X] = \frac{KM}{N}$.
2. $Var[X] = \frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}$.

Definition 3.3: Bernoulli Trial

A Bernoulli Trial is an experiment with two, and only two, possible outcomes. A random variable X has a Bernoulli(p) distribution if

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}, \qquad 0 \leq p \leq 1 \tag{3.3}$$

Definition 3.4: Binomial Distribution

Suppose that we have n Bernoulli Trials and consider a result of 1 as a success with probability p. Let Y denote the total number of successes in n trials. The probability distribution of y is therefore given by

$$P(Y = y|n, p) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y = 0, 1, 2, ..., n, \tag{3.4}$$

and Y is called a binomial(n, p) random variable.

Proposition 3.3: Properties of Binomial Distributed Variables

Suppose that $X \sim binomial(n, p)$, then

1. $E[X] = np$.
2. $Var[X] = np(1-p)$.
3. $M_X(t) = \left[pe^t + (1-p)\right]^n$

Theorem 3.1: Binomial Theorem

For any real numbers x and y and integer $n \geq 0$,

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}. \tag{3.5}$$

Definition 3.5: Poisson Distribution

The Poisson distribution is a widely applied discrete distribution and can serve as a model for a number of different types of experiments. For example, if we are modeling a phenomenon in which we are waiting for an occurrence (such as waiting for a bus), number of occurrences in a given time interval can sometimes be modeled by the Poisson distribution. The Poisson distribution has a single parameter λ, sometimes called the intensity parameter, where $\lambda > 0$. A random variable X, taking values in the nonnegative integers, has

a *Poisson*(λ) *distribution if*

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, ... \tag{3.6}$$

Proposition 3.4: Properties of Poisson Distributed Variables

Let $X \sim$ *Poisson*(λ)*, then*

1. $E[X] = \lambda$.
2. $Var[\lambda] = \lambda$.
3. $M_X(t) = e^{\lambda(e^t - 1)}$.

Definition 3.6: Negative Binomial(n, p) Distribution

In a sequence of independent Bernoulli(p) *trials, let the random variable* X *denote the trial at which the* r^{th} *success occurs, where* r *is a fixed integer. Then,*

$$P(X = x|r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, ... \tag{3.7}$$

and we say that X *has a negative binomial*(r, p) *distribution.*

Proposition 3.5: Properties of Negative Binomial Distributed Variables

Let X *have a negative binomial*(r, p) *distribution, then*

1. $E[X] = r\frac{1-p}{p}$.
2. $Var[X] = \frac{r(1-p)}{p^2}$

Definition 3.7: Geometric Distribution

The geometric distribution is one of the simplest of the waiting time distributions and is a special case of the negative binomial distribution. If we set $r = 1$*, we have*

$$P(X = x|p) = p(1-p)^{x-1}, \qquad x = 1, 2, ..., \tag{3.8}$$

which defines the PMF *of a geometric random variable* X *with success probability* p*.* X *can be interpreted as the trial at which the first success occurs, so we are "waiting for a success".*

Proposition 3.6: Properties of Geometrically Distributed Variables

Let X *have a geometric*(p) *distribution. Then,*

1. $E[X] = \frac{1}{p}$.
2. $Var[X] = \frac{1-p}{p^2}$.

3. *Geometric distribution has the property that if $s > t$, then*

$$P(X > s|X > t) = P(X > s - t) \tag{3.9}$$

3.2 Continuous Distributions

Definition 3.8: Uniform Distribution (Continuous)

The continuous uniform distribution is defined by spreading mass uniformly over an interval $[a, b]$. Its PDF is given by

$$f(x|a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

Proposition 3.7: Properties of Uniformly Distributed Variables

Let X be a uniformly distributed random variable $[a, b]$. Then

1. $E[X] = \frac{a+b}{2}$.
2. $Var[X] = \frac{(b-a)^2}{12}$

Definition 3.9: Gamma Function

Let $\Gamma : [0, \infty) \to \mathbb{R}$ defined by

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \tag{3.11}$$

Definition 3.10: Gamma Distribution (Continuous)

The gamma(α, β) family is described by

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \qquad 0 < x < \infty,\ \alpha > 0,\ \beta > 0. \tag{3.12}$$

The parameter α is known as the shape parameter, since it influences the peakedness of the distribution, while the parameter β is called the scale parameter. since most of its influence is on the spread of the distribution.

Proposition 3.8: Properties of Gamma Distributed Variables

Let $X \sim$ gamma(α, β). Then,

1. $E[X] = \alpha\beta$.
2. $Var[X] = \alpha\beta^2$.
3. $M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha, \qquad t < \frac{1}{\beta}$.

Definition 3.11: Chi-Squared Distribution

If we set $\alpha = p/2$, where p is an integer, and $\beta = 2$, then the Gamma PDF becomes

$$f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, \qquad 0 < x < \infty \tag{3.13}$$

which is the chi squared PDF with p degrees of freedom. We denote this distribution by χ^2_p. This distribution plays an important role in statistical inference, especially when sampling from a normal distribution. This will be dealt with in detail in Chapter 5.

Proposition 3.9: Properties of the Chi-Squared Distribution

Let $X \sim \chi^2_p$. Then,

1. $E[X] = p$.
2. $Var[X] = 2p$.
3. $M_X(t) = \left(\frac{1}{1-2t}\right)^{p/2}, \qquad t < 1/2.$

Definition 3.12: Exponential Distribution

Another special case of the gamma distribution is obtained when we set $\alpha = 1$. We then have

$$f(x|\beta) = \frac{1}{\beta} e^{-x/\beta}, \qquad 0 < x < \infty, \tag{3.14}$$

which we refer to as the exponential PDF with scale parameter β. The exponential distribution can be used to model lifetimes, analogous to the use of the geometric distribution in the discrete case.

Proposition 3.10: Properties of the Exponential Distributed Variables

Let $X \sim$ exponential(λ). Then

1. $E[X] = \lambda$.
2. $Var[X] = \lambda^2$.
3. $P(X > s|X > t) = P(X > s - t)$ *for* $s > t \geq 0$.

Definition 3.13: Weibull Distribution

If $X \sim$ exponential(β), then $Y = X^{1/\gamma}$ has a Weibull(γ, β) distribution,

$$f_Y(y|\gamma, \beta) = \frac{\gamma}{\beta} y^{\gamma-1} e^{-y^\gamma/\beta}, \qquad 0 < y < \infty,\ \gamma > 0,\ \beta > 0. \tag{3.15}$$

The Weibull distribution plays an extremely important role in the analysis of failure time data.

Definition 3.14: Normal / Gaussian Distribution

The normal / Gaussian distribution plays a central role in a large body of statistics. The normal distribution has two parameters, usually denoted by μ and σ^2, which are its mean and variance. The PDF of the normal distribution with mean μ and variance σ^2 (usually denoted by $n(\mu, \sigma^2)$ or $\mathcal{N}(\mu, \sigma^2)$) is given by

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \qquad -\infty < x < \infty \tag{3.16}$$

Definition 3.15: Standard Normal Distribution

If $X \sim n(\mu, \sigma^2)$, then the random variable $Z = (X - \mu)/\sigma$ has a $n(0, 1)$ distribution, also known as the standard normal.

Definition 3.16: Beta Function

The Beta Function integral representation is defined as follows:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx. \tag{3.17}$$

The beta function is related to the gamma function through the following identity:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{3.18}$$

Definition 3.17: Beta Distribution

The beta family of distributions is a continuous family on $(0, 1)$ indexed by two parameters. The beta(α, β) PDF is given by

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1,\ \alpha > 0,\ \beta > 0, \tag{3.19}$$

where $B(\alpha, \beta)$ denotes the beta function.

Proposition 3.11: Properties of the Beta Distributed Variables

Let $X \sim$ beta(α, β). Then

1. $E[X] = \frac{\alpha}{\alpha+\beta}$
2. $E[X^n] = \frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)\Gamma(\alpha)}$
3. $Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Definition 3.18: Cauchy Distribution

The Cauchy distribution is a symmetric, bell-shaped distribution on $(-\infty, \infty)$ with PDF

$$f(x|\theta) = \frac{1}{\pi} \frac{1}{1 + (x-\theta)^2}, \qquad -\infty < x < \infty,\ -\infty < \theta < \infty \tag{3.20}$$

A key property of the Cauchy Distribution is that $E[X] = \infty$ and consequently that all absolute moments do not exist.

Definition 3.19: Lognormal Distribution

If X is a random variable whose logarithm is normally distributed (that is, $log(X) \sim n(\mu, \sigma^2)$), then X has a lognormal distribution. The lognormal PDF is given by

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}e^{-(log(x)-\mu)^2/(2\sigma^2)}, \qquad 0 < x < \infty,\ -\infty < \mu < \infty,\ \sigma > 0 \tag{3.21}$$

Proposition 3.12: Properties of Lognormal Distributed Variables

Let $X \sim lognormal(\mu, \sigma^2)$. Then,

1. $E[X] = e^{\mu+(\sigma^2/2)}$.
2. $Var[X] = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$.

Definition 3.20: Double Exponential Distribution

The double exponential distribution is formed by reflecting the exponential distribution around its mean. The PDF is given by

$$f(x|\mu, \sigma^2) = \frac{1}{2\sigma}e^{-|x-\mu|/\sigma}, \qquad -\infty < x < \infty,\ -\infty < \mu < \infty,\ \sigma > 0 \tag{3.22}$$

Proposition 3.13: Properties of Double Exponential Distributed Variables

Let $X \sim double\text{-}exp(\mu, \sigma)$. Then,

1. $E[X] = \mu$.
2. $Var[X] = 2\sigma^2$.

3.3 Exponential Families

Many common families introduced in the previous section are exponential families. These include the continuous families - normal, gamma, and beta, and the discrete families - binomial, Poisson, and negative binomial.

Definition 3.21: Exponential Family

A family of PDFs or PMFs is called an exponential family if it can be expressed as

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right). \tag{3.23}$$

Here $h(x) \geq 0$ and $t_1(x), ..., t_k(x)$ are real-valued functions of the observation x (they cannot depend on θ), and $c(\theta) \geq 0$ and $w_1(\theta), ..., w_k(\theta)$ are real-valued functions of the possibly vector-valued parameter θ (they

cannot depend on x).

Theorem 3.2

If X is a random variable with PDF or PMF of the form (3.23), then

1. $$E\left[\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(X)\right] = -\frac{\partial}{\partial\theta_j}log\big(c(\theta)\big); \tag{3.24}$$

2. $$Var\left[\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(X)\right] = -\frac{\partial^2}{\partial\theta_j^2}log\big(c(\theta)\big) - E\left[\sum_{i=1}^{k}\frac{\partial^2 w_i(\theta)}{\partial\theta_j^2}t_i(X)\right] \tag{3.25}$$

Definition 3.22: Indicator Function

The indicator function of a set A, most often denoted by $I_A(x)$, is the function

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \tag{3.26}$$

An alternative notation is $I(x \in A)$.

Definition 3.23: Natural Parameter Space

An exponential family is sometimes reparameterized as

$$f(x|\eta) = h(x)c^*(\eta)exp\Big(\sum_{i=1}^{k}\eta_i t_i(x)\Big), \tag{3.27}$$

where $h(x)$ and $t_i(x)$ functions are the same as in the original parameterization (3.23). The set

$$\mathcal{H} = \left\{\eta = (\eta_1, ..., \eta_k) : \int_{-\infty}^{\infty} h(x)exp\Big(\sum_{i=1}^{k}\eta_i t_i(x)\Big)dx < \infty\right\} \tag{3.28}$$

is called the natural parameter space for the family. (The integral is replaced by a sum over the values of x for which $h(x) > 0$ if X is discrete.) For values of $\eta \in \mathcal{H}$, we must have

$$c^*(\eta) = \left[\int_{-\infty}^{\infty} h(x)exp\Big(\sum_{i=1}^{k}\eta_i t_i(x)\Big)dx\right]^{-1} \tag{3.29}$$

to ensure that $f(x|\eta)$ integrates to 1.

Definition 3.24: Curved Exponential Family

A curved exponential family is a family of densities of the form (3.23) for which the dimension of the vector θ is equal to $d < k$. If $d = k$, the family is a full exponential family.

3.4 Location and Scale Families

The three types of families are called *location families*, *scale families*, and *location-scale families*. Each of the families is constructed by specifying a single PDF, such as $f(x)$, called the *standard PDF* for the family. Then, all other PDFs in the family are generated by transforming the standard PDF in a prescribed way.

Theorem 3.3

Let $f(x)$ be any PDF and let μ and $\sigma > 0$ be any given constants. Then the function

$$g(x|\mu,\sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \tag{3.30}$$

is a PDF.

Definition 3.25: Location Family & Location Parameter

Let $f(x)$ be any PDF. Then the family of PDFs $f(x-\mu)$, indexed by the parameter $\mu, -\infty < \mu < \infty$, is called the location family with standard PDF $f(x)$ and μ is called the location parameter for the family.

Proposition 3.14

If X is a random variable with PDF $f(x-\mu)$, then X may be represented as $X = Z + \mu$, where Z is a random variable with PDF $f(z)$.

Definition 3.26: Scale Family & Scale Parameter

Let $f(x)$ be any PDF. Then for any $\sigma > 0$, the family of PDFs $(1/\sigma)f(x/\sigma)$, indexed by the parameter σ, is called the scale family with standard PDF $f(x)$ and σ is called the scale parameter of the family.

The effect of introducing the scale parameter σ is either to stretch ($\sigma > 1$) or to contract ($\sigma < 1$) the graph of $f(x)$ while still maintaining the same basic shape of the graph.

Definition 3.27: Location-Scale Family

Let $f(x)$ be any PDF. Then for any $\mu, -\infty < \mu < \infty$, and any $\sigma > 0$, the family of PDFs $(1/\sigma)f\big((x-\mu)/\sigma\big)$, indexed by the parameter (μ,σ), is called the location-scale family with standard PDF $f(x)$; μ is called the location parameter and σ is called the scale parameter.

Theorem 3.4

Let $f(\cdot)$ be any PDF. Let $\mu \in \mathbb{R}$, and let $\sigma > 0$. Then X is a random variable with PDF $(1/\sigma)f\big((x-\mu)/\sigma\big)$ if and only if there exists a random variable Z with PDF $f(z)$ and $X = \sigma Z + \mu$.

Theorem 3.5

Let Z be a random variable with PDF $f(z)$. Suppose that $E[Z]$ and $Var[Z]$ exist. If X is a random variable with PDF $(1/\sigma)f\big((x-\mu)/\sigma\big)$, then

$$E[X] = \sigma E[Z] + \mu \qquad \text{and} \qquad Var[X] = \sigma^2 Var[Z]. \tag{3.31}$$

In particular, if $E[Z] = 0$ and $Var[Z] = 1$, then $E[X] = \mu$ and $Var[X] = \sigma^2$.

3.5 Inequalities and Identities

Theorem 3.6: Chebychev's Inequality

Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$,

$$P(g(X) \geq r) \leq \frac{E[g(X)]}{r} \tag{3.32}$$

Theorem 3.7

let $X_{\alpha,\beta}$ denote a gamma(α, β) random variable with PDF $f(x|\alpha, \beta)$, where $\alpha > 1$. Then for any constants a and b,

$$P(a < X_{\alpha,\beta} < b) = \beta\big(f(a|\alpha,\beta) - f(b|\alpha,\beta)\big) + P(a < X_{\alpha-1,\beta} < b) \tag{3.33}$$

Lemma 3.1: Stein's Lemma

Let $X \sim n(\theta, \sigma^2)$, and let g be a differentiable function satisfying $E\big[|g'(X)|\big] < \infty$. Then

$$E[g(X)(X - \theta)] = \sigma^2 E[g'(X)] \tag{3.34}$$

As an example, Stein's Lemma makes calculations of higher-order moments for normally distributed variables quite easy.

Theorem 3.8

Let χ^2_p denote a chi squared random variable with p degrees of freedom. For any function $h(x)$,

$$E[h(\chi^2_p)] = pE\left[\frac{h(\chi^2_{p+2})}{\chi^2_{p+2}}\right], \tag{3.35}$$

provided the expectations exist.

Theorem 3.9: Hwang's Theorem

Let $g(x)$ be a function satisfying $-\infty < E[g(X)] < \infty$ and $-\infty < g(-1) < \infty$. Then

a. *If $X \sim$ Poisson(λ),*

$$E[\lambda g(X)] = E[Xg(X-1)]. \tag{3.36}$$

b. *If $X \sim$ negative binomial(r, p),*

$$E[(1-p)g(X)] = E\left[\frac{X}{r+X-1}g(X-1)\right] \tag{3.37}$$

3.6 Miscellanea

Theorem 3.10

For each $t \geq 0$, let N_t be an integer-valued random variable with the following properties. (Think of N_t as denoting the number of arrivals in the time period from time 0 to time t.)

i) $N_0 = 0$ *(Start with no arrivals)*

ii) If $s < t$, *then* N_s *and* $N_t - N - s$ *are independent (Arrivals in Disjoint Time periods are independent)*

iii) N_s *and* $N_{t+s} - N_t$ *are identically distributed. (Number of arrivals depends only on period length)*

iv) $\lim_{t\to 0} \frac{P(N_t=1}{t} = \lambda$ *(Arrival Probability Proportional to period length, if length is small)*

v) $\lim_{t\to 0} \frac{P(N_t>0)}{t} = 0$ *(No simultaneous arrivals)*

If $i - v$ hold, then for any integer n

$$P(N_t = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \tag{3.38}$$

that is, $N_t \sim Poisson(\lambda t)$.

Theorem 3.11

Let $X_1, ..., X_n$ be identically distributed random variables with $\mu := E[X]$ and $\sigma^2 = Var[X]$. Then, Chebychev's Inequality states

$$P(|\bar{X}_n - \mu| \geq k\sigma) \leq \frac{1}{nk^2}, \tag{3.39}$$

where $\bar{X}_n$ is the mean of the random variables. If $0 < \sigma < \infty$, then

a. *If $n = 1$, the inequality is attainable for $k \geq 1$ and unattainable for $0 < k < 1$.*

b. *If $n = 2$, the inequality is attainable if and only if $k = 1$.*

c. *If $n \geq 3$, the inequality is not attainable.*

Lemma 3.2: Markov's Inequality

If $P(Y \geq 0) = 1$ and $P(Y = 0) < 1$, then for any $r > 0$

$$P(Y \leq r) \leq \frac{E[Y]}{r} \tag{3.40}$$

with equality if and only if $P(Y = r) = p = 1 - P(Y = 0)$, $0 < p \leq 1$.

Definition 3.28: Mode

1. *If X is a discrete random variable with PMF f, then the mode of X is the value m satisfying $m := \text{argmax}_x f(x)$. If one such value exists, then the distribution is said to be unimodal, otherwise it is considered multimodal.*
2. *If X is a continuous random variable with PDF f, then the modes of X are the values m such that $f(m)$ is a local maximum. For a distribution to be unimodal, only one such mode can exist, otherwise it is considered to be multimodal.*

Theorem 3.12: Gauss Inequality

Let $X \sim f$, where f is unimodal with mode ν, and define $\tau^2 = E[(X-\nu)^2]$. Then

$$P(|X-\nu| > \epsilon) \leq \begin{cases} \frac{4\tau^2}{9\epsilon^2} & \text{for all } \epsilon \geq \sqrt{4/3}\tau \\ 1 - \frac{\epsilon}{\sqrt{3}\tau} & \text{for all } \epsilon \leq \sqrt{4/3}\tau \end{cases} \tag{3.41}$$

Theorem 3.13: Vysochanskii-Petunin Inequality

Let $X \sim f$, where f is unimodal, and define $\xi^2 = E[(X-\alpha)^2]$ for an arbitrary point α. Then

$$P(|X-\alpha| > \epsilon) \leq \begin{cases} \frac{4\xi^2}{9\epsilon^2} & \text{for all } \epsilon \geq \sqrt{8/3}\xi \\ \frac{4\xi^2}{9\epsilon^2} - \frac{1}{3} & \text{for all } \epsilon \leq \sqrt{8/3}\xi \end{cases} \tag{3.42}$$

Chapter 4

Multiple Random Variables

4.1 Joint and Marginal Distributions

In the previous chapters, we discussed probability models and computation of probability for events involving only one random variable. These are called *univariate models*. In this chapter, we discuss probability models that involve more than one random variable-naturally enough, called *multivariate models*.

Definition 4.1: Random Vector

An n-dimensional random vector is a function from a space space S into $\mathbb{R}^n$, n-dimensional Euclidean space.

Definition 4.2: Discrete Random Vector

A random vector that has a countable number of possible values is referred to as a discrete random vector.

Definition 4.3: Joint Probability Mass Function (PMF)

Let (X, Y) be a discrete bivariate random vector. Then the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = P(X = x, Y = y)$ is called the joint probability mass function or joint PMF of (X, Y). If it is necessary to stress the fact that f is the joint PMF of the vector (X, Y) rather than some other vector, the notation $f_{X,Y}(x, y)$ will be used.

Notation 4.1: Probability of Event

The joint PMF can be used to compute the probability of any event defined in terms of (X, Y). Let $A \subset \mathbb{R}^2$, then

$$P\big((X, Y) \in A\big) = \sum_{(x,y)\in A} f(x, y) \tag{4.1}$$

Definition 4.4: Expectation of Discrete Bivariate Model

Let $g : \mathbb{R}^2 \to \mathbb{R}$ and (X, Y) be a random vector with joint PMF $f(x, y)$. Then

$$E[g(X, Y)] = \sum_{(x,y)\in\mathbb{R}^2} g(x, y)f(x, y) \tag{4.2}$$

Theorem 4.1: Marginal PMFs

Let (X,Y) be a discrete bivariate random vector with joint PMF $f_{X,Y}(x,y)$. Then the marginal PMFs of X and Y, $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$, are given by

$$f_X(x) = \sum_{y\in\mathbb{R}} f_{X,Y}(x,y) \qquad \text{and} \qquad f_Y(y) = \sum_{x\in\mathbb{R}} f_{X,Y}(x,y) \tag{4.3}$$

Definition 4.5: Joint Probability Density Function (Joint PDF)

A function $f : \mathbb{R}^2 \to \mathbb{R}$ is called a joint probability density function or joint PDF of the continuous bivariate random vector (X,Y) if, for every $A \subset \mathbb{R}^2$,

$$P\big((X,Y) \in A\big) = \int\int_A f(x,y)\ dxdy \tag{4.4}$$

Definition 4.6: Expectation of Continuous Bivariate Model

Let $g : \mathbb{R}^2 \to \mathbb{R}$ and (X,Y) be a random vector with joint PDF $f(x,y)$. Then the expected value of $g(X,Y)$ is defined to be

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\ dxdy \tag{4.5}$$

Definition 4.7: Marginal Probability Density Functions (Marginal PDF)

Let (X,Y) be a random vector with joint PDF $f(x,y)$. Then the marginal PDF of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy, \qquad -\infty < x < \infty, \tag{4.6}$$

and the marginal PDF of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx, \qquad -\infty < y < \infty, \tag{4.7}$$

Proposition 4.1: Joint PDF Properties

Any function $f : \mathbb{R}^2 \to \mathbb{R}$ satisfying

1. $f(x,y) \geq 0\ \forall (x,y) \in \mathbb{R}^2$,
2. $1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy$

is the joint PDF of some continuous bivariate random vector (X,Y).

Definition 4.8: Joint Cumulative Probability Distribution (Joint CDF)

Let (X,Y) be a random vector. The joint CDF is the function $F(x,y)$ defined by

$$F(x,y) = P(X \leq x, Y \leq y) \tag{4.8}$$

Corollary 4.1

Let (X, Y) be a continuous random vector with PDF $f(s,t)$. The joint CDF is then defined as

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(s,t)\ dtds. \tag{4.9}$$

Hence, at the continuity points of $f(x,y)$ we have the relationship

$$\frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y) \tag{4.10}$$

4.2 Conditional Distributions and Independence

Definition 4.9: Conditional PMF

Let (X, Y) be a discrete bivariate random vector with joint PMF $f(x,y)$ and marginal PMFs $f_X(x)$ and $f_Y(y)$. For any x such that $P(X = x) = f_X(x) > 0$ the conditional PMF of Y given that $X = x$ is the function of y denoted by $f(y|x)$ and defined by

$$f(y|x) = P(Y = y|X = x) = \frac{f(x,y)}{f_X(x)}. \tag{4.11}$$

Similarly, for any y such that $P(Y = y) = f_Y(y) > 0$, the conditional PMF of X given that $Y = y$ is the function of x denoted by $f(x|y)$ and defined by

$$f(x|y) = P(X = x|Y = y) = \frac{f(x,y)}{f_Y(y)} \tag{4.12}$$

Definition 4.10: Conditional PDF

Let (X, Y) be a continuous bivariate random vector with joint PDF $f(x,y)$ and marginal PDFs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$ the conditional PDF of Y given that $X = x$ is the function of y denoted by $f(y|x)$ and defined by

$$f(y|x) = P(Y = y|X = x) = \frac{f(x,y)}{f_X(x)}. \tag{4.13}$$

Similarly, for any y such that $f_Y(y) > 0$, the conditional PDF of X given that $Y = y$ is the function of x denoted by $f(x|y)$ and defined by

$$f(x|y) = P(X = x|Y = y) = \frac{f(x,y)}{f_Y(y)} \tag{4.14}$$

Definition 4.11: Conditional Expectation

Let (X, Y) be a random vector. Let $g(Y)$ be a function of Y, then the conditional expected value of $g(Y)$ given that $X = x$ is denoted by $E[g(Y)|x]$ and is given by

$$E[g(Y)|x] = \sum_y g(y)f(y|x) \qquad \text{and} \qquad E[g(Y)|x] = \int_{-\infty}^{\infty} g(y)f(y|x)\ dy \tag{4.15}$$

for the discrete and continuous cases, respectively. The conditional expectation shares all the same properties with the regular expectation as seen in Theorem 2.5.

Definition 4.12: Conditional Variance

The variance of the probability distribution described by $f(y|x)$ is called the conditional variance of Y given $X = x$. Using the notation $Var[Y|x]$ for this, we have

$$Var[Y|x] = E[Y^2|x] - \big(E[Y|x]\big)^2 \tag{4.16}$$

Definition 4.13: Independent Random Variables

Let (X, Y) be a bivariate random vector with joint PDF or PMF $f(x, y)$ and marginal PDFs or PMFs $f_X(x)$ and $f_Y(y)$. Then X and Y are called independent random variable if, for every $(x, y) \in \mathbb{R}^2$,

$$f(x, y) = f_X(x) f_Y(y) \tag{4.17}$$

Lemma 4.1

Let (X, Y) be a bivariate random vector with joint PDF or PMF $f(x, y)$. Then X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $(x, y) \in \mathbb{R}^2$,

$$f(x, y) = g(x)h(y). \tag{4.18}$$

Theorem 4.2

Let X and Y be independent random variables.

- **a.** *For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$; that is, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events.*
- **b.** *Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then*

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)] \tag{4.19}$$

Theorem 4.3

Let X and Y be independent random variables with moment generating functions $M_X(t)$ and $M_Y(t)$. Then the moment generating function of the random variable $Z = X + Y$ is given by

$$M_Z(t) = M_X(t)M_Y(t) \tag{4.20}$$

Theorem 4.4

Let $X \sim n(\mu, \sigma^2)$ and $Y \sim n(\gamma, \tau^2)$ be independent normal random variables. Then the random variable $Z = X + Y$ has a $n(\mu + \gamma, \sigma^2 + \tau^2)$ distribution.

4.3 Bivariate Transformations

Bivariate Transformation for Discrete Random Vector

Let (X,Y) be a bivariate random vector with a known probability distribution. Now consider a new bivariate random vector (U,V) defined by $U = g_1(X,Y)$ and $V = g_2(X,Y)$, where $g_1(x,y)$ and $g_2(x,y)$ are some specified functions. If $B \subset \mathbb{R}^2$, then $(U,V) \in B$ if and only if $(X,Y) \in A$, where $A = \{(x,y) : \big(g_1(x,y), g_2(x,y)\big) \in B\}$. Thus $P\big((U,V) \in B\big) = P\big((X,Y) \in A\big)$. If (X,Y) is a discrete bivariate random vector, then we define

$$\mathcal{A} := \{(x,y) \in \mathbb{R}^2 : f(x,y) > 0\} \tag{4.21}$$

and

$$\mathcal{B} := \{(u,v) : u = g_1(x,y) \text{ and } v = g_2(x,y) \text{ for some} (x,y) \in \mathcal{A}\}. \tag{4.22}$$

We now define the set

$$A_{uv} := \{(x,y) \in \mathcal{A} : g_1(x,y) = u \text{ and } g_2(x,y) = v\} \tag{4.23}$$

Then, the joint PMF of (U,V), given by $f_{U,V}(u,v)$ can be computed from the joint PMF of (X,Y) by

$$f_{U,V}(u,v) = P(U = u, V = v) = P\big((X,Y) \in A_{uv}\big) = \sum_{(x,y)\in A_{uv}} f_{X,Y}(x,y) \tag{4.24}$$

Theorem 4.5

If $X \sim$ Poisson(θ) and $Y \sim$ Poisson(λ) and X and Y are independent, then $X + Y \sim$Poisson$(\theta + \lambda)$.

Bivariate Transformation for Continuous Random Vector

If (X,Y) is a continuous random vector with joint PDF $f_{X,Y}(x,y)$, the the joint PDF of (U,V) can be expressed in terms of $f_{X,Y}(x,y)$ in a manner analogous to (2.8). We once again define

$$\mathcal{A} := \{(x,y) : f_{X,Y}(x,y) > 0\} \tag{4.25}$$

and

$$\mathcal{B} := \{(u,v) : u = g_1(x,y) \text{ and } v = g_2(x,y) \text{ for some } (x,y) \in \mathcal{A}\} \tag{4.26}$$

Suppose that the transformation $u = g_1(x,y)$ and $v = g_2(x,y)$ is one-to-one, then by definition of $\mathcal{B}$, this turns $(x,y) \mapsto (u,v)$ into a bijection. Hence, there is only one $(x,y) \in \mathcal{A}$ such that $(u,v) = (g_1(x,y), g_2(x,y))$. For such a relationship, we can define the inverse transformations by

$$x = h_1(u,v), \qquad y = h_2(u,v). \tag{4.27}$$

Let J be the Jacobian of the transformation, recalling that it is the determinant of a matrix of partial derivatives, defined by

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}. \tag{4.28}$$

Assuming that J is not identically 0 on $\mathcal{B}$, then the joint PDF of (U,V) is 0 outside the set $\mathcal{B}$ and on the set $\mathcal{B}$ is given by

$$f_{U,V}(u,v) = f_{X,Y}\big(h_1(u,v), h_2(u,v)\big)|J|, \tag{4.29}$$

where $|J|$ is the absolute value of J.

Theorem 4.6: Independence of Transformed Variables

Let X and Y be independent random variables. let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then the random variables $U = g(X)$ and $V = h(Y)$ are independent.

4.4 Hierarchical Models and Mixture Distributions

Theorem 4.7

If X and Y are any two random variables, then

$$E[X] = E[E[X|Y]], \tag{4.30}$$

provided that the expectations exist.

Definition 4.14: Mixture Distribution

A random variable X is said to have a mixture distribution if the distribution of X depends on a quantity that also has a distribution.

Definition 4.15: Noncentral Chi Squared Distribution

A distribution that often occurs in statistics is the noncentral chi squared distribution. With p degrees of freedom and noncentrality parameter λ, the PDF is given by

$$f(x|\lambda, p) = \sum_{k=0}^{\infty} \frac{x^{p/2+k-1}e^{-x/2}}{\Gamma(p/2+k)2^{p/2+k}} \frac{\lambda^k e^{-\lambda}}{k!} \tag{4.31}$$

Theorem 4.8: Conditional Variance Identity

For any two random variables X and Y,

$$Var[X] = E\big[Var(X|Y)\big] + Var\big(E[X|Y]\big), \tag{4.32}$$

provided that the expectations exist.

4.5 Covariance and Correlation

In this section, we'll be frequently referring to the mean and variance of X and the mean and variance of Y. We will use the notation $E[X] = \mu_X$, $E[Y] = \mu_Y$, $Var[X] = \sigma_X^2$ and $Var[Y] = \sigma_Y^2$. We will assume that $0 < \sigma_X^2 < \infty$ and $0 < \sigma_Y^2 < \infty$.

Definition 4.16: Covariance

The covariance of X and Y is the value defined by

$$Cov(X, Y) = E\big[(X - \mu_X)(Y - \mu_Y)\big] \tag{4.33}$$

Definition 4.17: Correlation

The correlation of X and Y is the value defined by

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}. \tag{4.34}$$

The value ρ_{XY} is also called the correlation coefficient.

Proposition 4.2: Covariance Identity

For any random variables X and Y,

$$Cov(X,Y) = E[XY] - \mu_X \mu_Y \tag{4.35}$$

Proposition 4.3: Independence and Zero Covariance

If X and Y are independent random variables, then $Cov(X,Y) = 0$ and $\rho_{XY} = 0$.

Theorem 4.9

If X and Y are any two random variables and $a, b \in \mathbb{R}$, then

$$Var[aX + bY] = a^2 Var[X] + b^2 Var[Y] + 2ab Cov(X,Y) \tag{4.36}$$

Theorem 4.10

For any random variables X and Y,

- **a.** $-1 \leq \rho_{XY} \leq 1$.
- **b.** $|\rho_{XY}| = 1$ *if and only if there exist numbers $a \neq 0$ and b such that $P(Y = aX + b) = 1$. If $\rho_{XY} = 1$, then $a > 0$, and if $\rho_{XY} = -1$, then $a < 0$.*

Definition 4.18: Bivariate Normal PDF

Let $-\infty < \mu_X < \infty$, $-\infty < \mu_Y < \infty$, $0 < \sigma_X$, $0 < \sigma_Y$, and $-1 < \rho < 1$ be five real numbers. The bivariate normal PDF with means μ_X and μ_Y. The variances σ_X^2 and σ_Y^2, and correlation ρ is the bivariate PDF given by

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 \right] \right\} \tag{4.37}$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$.

Proposition 4.4: Properties of Bivariate Normal Distribution

Consider the bivariate normal PDF defined in Def. 4.18. The properties of this distribution include:

- **a.** *The marginal distribution of X is $n(\mu_X, \sigma_X^2)$.*
- **b.** *The marginal distribution of Y is $n(\mu_Y, \sigma_Y^2)$.*

c. *The correlation between X and Y is $\rho_{XY} = \rho$.*

d. *For any constants a and b, the distribution of $aX + bY$ is $n(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$.*

4.6 Multivariate Distributions

Notation 4.2: Random Vector

In this section we consider a multivariate random vector $(X_1, ..., X_n)$. We will use boldface letters to denote multiple variates. Thus, we write $\mathbf{X}$ to denote the random variables $X_1, ..., X_n$ and $\mathbf{x}$ to denote the sample $x_1, ..., x_n$.

Notation 4.3: Discrete Random Vector Distribution

The random vector $\mathbf{X} = (X_1, ..., X_n)$ has a "sample space" that is a subset of $\mathbb{R}^n$. If $(X_1, ..., X_n)$ is a discrete random vector, then the sample space is countable. Consequently, the joint PMF of $(X_1, ..., X_n)$ is the function defined by

$$f(\mathbf{x}) = f(x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n) \tag{4.38}$$

for each $(x_1, ..., x_n) \in \mathbb{R}^n$. Then for any $A \subset \mathbb{R}^n$,

$$P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x}) \tag{4.39}$$

Notation 4.4: Continuous Random Vector Distribution

The random vector $\mathbf{X} = (X_1, ..., X_n)$ has a "sample space" that is a subset of $\mathbb{R}^n$. If $(X_1, ..., X_n)$ is a continuous random vector, then the sample space is uncountable. Consequently, the joint PDF of $(X_1, ..., X_n)$ is a function $f(x_1, ..., x_n)$ that satisfies

$$P(\mathbf{X} \in A) = \int \cdots \int_A f(\mathbf{x}) d\mathbf{x} = \int \cdots \int_A f(x_1, ..., x_n)\ dx_1 \cdots dx_n \tag{4.40}$$

These integrals are n-fold integrals with limits of integration set so that the integration is over all points $\mathbf{x} \in A$.

Definition 4.19: Expected Value of Random Vector

Let $g(\mathbf{x}) = g(x_1, ..., x_n)$ be a real-valued function defined on the "sample space" of $\mathbf{X}$. Let f denote the PMF / PDF of $\mathbf{X}$ if $\mathbf{X}$ is discrete / continuous. Then $g(\mathbf{X})$ is a random variable and the expected value of $g(\mathbf{X})$ is

$$E[g(\mathbf{X})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x})\ d\mathbf{x} \qquad \text{and} \qquad E[g(\mathbf{X})] = \sum_{x \in \mathbb{R}^n} g(\mathbf{x}) f(\mathbf{x}) \tag{4.41}$$

in the continuous and discrete cases, respectively.

Definition 4.20: Marginal PDF / PMF

The marginal PDF or PMF of any subset of the coordinates of $(X_1, ..., X_n)$ can be computed by integrating or summing the joint PDF) or PMF over all possible values of the other coordinates. For instance the marginal distribution of the first k coordinates $(X_1, ..., X_k)$, is given by the PDF

$$f(x_1, ..., x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, ..., x_n)\, dx_{k+1} \cdots dx_n \tag{4.42}$$

or

$$f(x_1, ..., x_k) = \sum_{(x_{k+1}, ..., x_n) \in \mathbb{R}^{n-k}} f(x_1, ..., x_n) \tag{4.43}$$

for every $(x_1, ..., x_k) \in \mathbb{R}^k$.

Definition 4.21: Conditional PDF / PMF

The conditional PDF or PMF of a subset of the coordinates $(X_1, ..., X_n)$ given the values of the remaining coordinates is obtained by dividing the joint PDF or PMF by the marginal PDF or PMF of the remaining coordinates. If $f(x_1, ..., x_k) > 0$, the conditional PDF or PMF of $(X_{k+1}, ..., X_n)$ given $X_1 = x_1, ..., X_k = x_k$ is the function of $(x_{k+1}, ..., x_n)$ defined by

$$f_{X_{k+1}, .., X_n}(x_{k+1}, ..., x_n | x_1, ..., x_k) = \frac{f_{X_1, .., X_n}(x_1, ..., x_n)}{f_{X_1, .., X_k}(x_1 ..., x_k)} \tag{4.44}$$

Definition 4.22: Multinomial Distribution

Let n and m be positive integers and let $p_1, ..., p_n$ be numbers satisfying $0 \leq p_i \leq 1$, $i = 1, ..., n$ and $\sum_{i=1}^n p_i = 1$. Then the random vector $(X_1, ..., X_n)$ has a multinomial distribution with m trials and cell probabilities $p_1, ..., p_n$ if the joint PMF of $(X_1, ..., X_n)$ is

$$f(x_1, ..., x_n) = \frac{m!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n} = m! \prod_{i=1}^n \frac{p_i^{x_i}}{x_i!} \tag{4.45}$$

on the set of $(x_1, ..., x_n)$ such that each x_i is a nonnegative integer and $\sum_{i=1}^n x_i = m$.

Definition 4.23: Multinomial Coefficient

The factor $m!/(x_1! x_2! \cdots x_n!)$ is called a multinomial coefficient. It is the number of ways that m objects can be divided into n groups with x_1 in the first group, x_2 in the second group, ... and x_n in the n^{th} group.

Theorem 4.11: Multinomial Theorem

Let m and n be positive integers. Let $\mathcal{A}$ be the set of vectors $\mathbf{x} = (x_1, ..., x_n)$ such that each x_i is a nonnegative integer and $\sum_{i=1}^n x_i = m$. Then for any real numbers $p_1, ..., p_n$,

$$(p_1 + \cdots + p_n)^m = \sum_{\mathbf{x} \in \mathcal{A}} \frac{m!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n} \tag{4.46}$$

Definition 4.24: Mutually Independent Random Vectors

Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be random vectors with joint PDF or PMF $f(\mathbf{x}_1, ..., \mathbf{x}_n)$. Let $f_{\mathbf{X}_i}(\mathbf{x}_i)$ denote the marginal PDF or PMF of $\mathbf{X}_i$. Then $\mathbf{X}_1, ..., \mathbf{X}_n$ are called mutually independent random vectors if, for every $(\mathbf{x}_1, ..., \mathbf{x}_n)$,

$$f(\mathbf{x}_1, ..., \mathbf{x}_n) = f_{\mathbf{X}_1}(\mathbf{x}_1) \cdots f_{\mathbf{X}_n}(\mathbf{x}_n) = \prod_{i=1}^{n} f_{\mathbf{X}_i}(\mathbf{x}_i). \tag{4.47}$$

If the X_i's are all one-dimensional, then $X_1, ..., X_n$ are called mutually independent random variables. Mutual independence implies that any pair X_i and X_j are pairwise independent.

Theorem 4.12

Let $X_1, ..., X_n$ be mutually independent random variables. Let $g_1, ..., g_n$ be real-valued functions such that $g_i(x_i)$ is a function only of x_i, $i = 1, ..., n$. Then

$$E\big[g_1(X_1) \cdots g_n(X_n)\big] = E[g_1(X_1)] \cdots E[g_n(X_n)] \tag{4.48}$$

Theorem 4.13

Let $X_1, ..., X_n$ be mutually independent random variables with MGFs $M_{X_1}(t), ..., M_{X_n}(t)$. Let $Z = X_1 + \cdots + X_n$. Then the MGF of Z is

$$M_Z(t) = M_{X_1}(t) \cdots M_{X_n}(t). \tag{4.49}$$

In particular, if $X_1, ..., X_n$ all have the same distribution with MGF $M_X(t)$, then

$$M_Z(t) = \big(M_X(t)\big)^n \tag{4.50}$$

Corollary 4.2

Let $X_1, .., X_n$ be mutually independent random variables with MGFs $M_{X_1}(t), ..., M_{X_n}(t)$. Let $a_1, ..., a_n$ and $b_1, ..., b_n$ be fixed constants. Let $Z = \sum_{i=1}^{n} a_i X_i + b_i$. Then the MGF of Z is

$$M_Z(t) = e^{t \sum b_i} M_{X_1}(at) \cdots M_{X_n}(a_n t) \tag{4.51}$$

Corollary 4.3

Let $X_1, ..., X_n$ be mutually independent random variables with $X_i \sim n(\mu_i, \sigma_i^2)$. Let $a_1, ..., a_n$ and $b_1, ..., b_n$ be fixed constants. Then

$$Z = \sum_{i=1}^{n} (a_i X_i + b_i) \sim n\left(\sum_{i=1}^{n} (a_i \mu_i + b_i), \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right) \tag{4.52}$$

Theorem 4.14

Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be random vectors. Then $\mathbf{X}_1, ..., \mathbf{X}_n$ are mutually independent random vectors if and only if there exist functions $g_i(\mathbf{x}_i), i = 1, ..., n$, such that the joint PDF or PMF of $(\mathbf{X}_1, ..., \mathbf{X}_n)$ can be written as

$$f(\mathbf{x}_1, , ..., \mathbf{x}_n) = g_1(\mathbf{x}_1) \cdots g_n(\mathbf{x}_n) \tag{4.53}$$

Theorem 4.15

Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be independent random vectors. Let $g_i(\mathbf{x}_i)$ be a function only of $\mathbf{x}_i$, $i = 1, ..., n$. Then the random variable $U_i = g_i(\mathbf{X}_i), i = 1, ..., n$ are mutually independent.

Transformation of Continuous Random Vector

Let $(X_1, ..., X_n)$ be a random vector with PDF $f_{\mathbf{X}}(x_1, ..., x_n)$. Let $\mathcal{A} = \{\mathbf{x} : f_{\mathbf{x}}(\mathbf{x}) > 0\}$. Consider a new random vector $(U_1, ..., U_n)$, defined by $U_1 = g_1(X_1, ..., X_n), U_2 = g_2(X_1, ..., X_n), ..., U_n = g_n(X_1, ..., X_n)$. Suppose that $A_0, A_1, ..., A_k$ form a partition of $\mathcal{A}$ with these properties:

1. *The set A_0 which may be empty, satisfies $P((X_1, ..., X_n) \in A_0) = 0$.*
2. *The transformation $(U_1, ..., U_n) = (g_1(\mathbf{X}), ..., g_n(\mathbf{X}))$ is a one-to-one transformation from A_i onto $\mathcal{B}$ for each $i = 1, 2, ..., k$.*

Then, for each i, the inverse functions from $\mathcal{B}$ to A_i can be found. We denote the i^{th} inverse by

$$x_1 = h_{1i}(u_1, ..., u_n),\ x_2 = h_{2i}(u_1, ..., u_n), ...,\ x_n = h_{ni}(u_1, ..., u_n) \tag{4.54}$$

This i^{th} inverse gives for $(u_1, ..., u_n) \in \mathcal{B}$, the unique $(x_1, ..., x_n) \in A_i$ such that

$$(u_1, ..., u_n) = \big(g_1(x_1, ..., x_n), ..., g_n(x_1, ..., x_n)\big). \tag{4.55}$$

Let J_i denote the Jacobian computed from the i^{th} inverse. That is,

$$J_i = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \dots & \frac{\partial x_1}{\partial u_n} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} & \dots & \frac{\partial x_2}{\partial u_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \frac{\partial x_n}{\partial u_2} & \dots & \frac{\partial x_n}{\partial u_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial h_{1i}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{1i}(\mathbf{u})}{\partial u_2} & \dots & \frac{\partial h_{1i}(\mathbf{u})}{\partial u_n} \\ \frac{\partial h_{2i}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{2i}(\mathbf{u})}{\partial u_2} & \dots & \frac{\partial h_{2i}(\mathbf{u})}{\partial u_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial h_{ni}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{ni}(\mathbf{u})}{\partial u_2} & \dots & \frac{\partial h_{ni}(\mathbf{u})}{\partial u_n} \end{vmatrix}, \tag{4.56}$$

the determinant of an $n \times n$ matrix. Assuming that these Jacobians do not vanish identically on $\mathcal{B}$, we have the following representation of the joint PDF, $f_{\mathbf{U}}(u_1, ..., u_n)$, for $\mathbf{u} \in \mathcal{B}$,

$$f_{\mathbf{U}}(u_1, ..., u_n) = \sum_{i=1}^{k} f_{\mathbf{X}}\big(h_{1i}(u_1, ..., u_n), ..., h_{ni}(u_1, ..., u_n)\big)|J_i| \tag{4.57}$$

4.7 Inequalities

4.7.1 Numerical Inequalities

Lemma 4.2

Let $a, b > 0$, and let p and q be any positive numbers (necessarily greater than 1) satisfying

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{4.58}$$

Then

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab \tag{4.59}$$

with equality if and only if $a^p = b^q$.

Theorem 4.16: Holders Inequality

Let X and Y be any two random variables and let p and q satisfy (4.58). Then

$$|E[XY]| \leq E[|XY|] \leq \big(E[|X|^p]\big)^{1/p}\big(E[|Y|^q]\big)^{1/q}. \tag{4.60}$$

Corollary 4.4: Cauchy-Schwarz Inequality

For any two random variables X and Y,

$$|E[XY]| \leq E[|XY|] \leq (E[|X|^2])^{1/2}(E[|Y|^2])^{1/2} \tag{4.61}$$

Theorem 4.17: Minkowski's Inequality

Let X and Y be any two random variables. Then for $1 \leq p < \infty$,

$$\big(E[|X+Y|^p]\big)^{1/p} \leq \big(E[|X|^p]\big)^{1/p} + \big(E[|Y|^p]\big)^{1/p} \tag{4.62}$$

4.7.2 Functional Inequalities

Definition 4.25: Convex

A function $g(x)$ is said to be convex if $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$, for all x and y, and $0 < \lambda < 1$. The function $g(x)$ is concave if $-g(x)$ is convex.

Theorem 4.18: Jensen's Inequality

For any random variable X, if $g(x)$ is a convex function, then

$$E[g(X)] \leq g(E[X]), \tag{4.63}$$

where equality holds if and only if, for every line $a + bx$ that is tangent to $g(x)$ at $x = E[X]$, $P\big(g(X) = a + bX\big) = 1$.

Theorem 4.19: Covariance Inequality

Let X be any random variable and $g(x)$ and $h(x)$ any functions such that $E[g(X)]$, $E[h(X)]$, and $E[g(X)h(X)]$ exist.

a. *If $g(x)$ is a nondecreasing function and $h(x)$ is a nonincreasing function, then*

$$E[g(X)h(X)] \leq E[g(X)]E[h(X)] \tag{4.64}$$

b. *If $g(x)$ and $h(x)$ are either both nondecreasing or both nonincreasing, then*

$$E[g(X)h(X)] \leq E[g(X)]E[h(X)] \tag{4.65}$$

Chapter 5

Properties of a Random Sample

5.1 Basic Concepts of Random Samples

Definition 5.1: Random Sample

The random variables $X_1, ..., X_n$ are called a random sample of size n from the population $f(x)$ if $X_1, ..., X_n$ are mutually independent random variables and the marginal PDF or PMF of each X_i is the same function $f(x)$. Alternatively, $X_1, ..., X_n$ are called independent and identically distributed random variables with PDF or PMF $f(x)$. This is commonly abbreviated to iid random variables.

The random sampling model describes a type of experimental situation in which the variable of interest has a probability distribution described by $f(x)$. This random sampling model is sometimes called sampling from an *infinite* population. The assumption of independence in random sampling implies that the probability distribution for X_2 is unaffected by the fact that $X_1 = x_1$ was observed first. When sampling is from a *finite* population, Definition 5.1 may or may not be relevant depending on how the data collection is done.

Proposition 5.1: Joint PDF of Random Sample

Let $X_1, ..., X_n$ be a random sample from an underlying distribution f. Their joint PDF or PMF is given by

$$f_{X_1,..,X_n}(x_1, ..., x_n) = f(x_1)f(x_2)\cdots f(x_n) = \prod_{i=1}^{n} f(x_i) \tag{5.1}$$

In particular, if the population PDF or PMF is a member of a parametric family, with PDF or PMF given by $f(x|\theta)$, then the joint PDF or PMF is

$$f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{5.2}$$

Definition 5.2: Sampling With Replacement

Suppose one has a finite population of size N for which each of the N values is equally likely (probability = $1/N$) of being chosen. The process of drawing from the N values is repeated n times, yielding the sample $X_1, ..., X_n$. We would call this sampling with replacement if the value chosen at any stage is replaced in the population and is available for choice again at the next stage.

Definition 5.3: Sampling Without Replacement / Simple Random Sampling

A second method for drawing a random sample from a finite population is called sampling without replacement. Suppose one has a finite population of size N where a value is chosen from $\{x_1, ..., x_N\}$ in such a way that each of the N values has probability $1/N$ of being chosen. This value is recorded as $X_1 = x_{\sigma(1)}$ and then a second value is chosen from the remaining $N-1$ values. Each of the remaining $N-1$ values has probability $1/(N-1)$ of being chosen. The choice of the remaining values continues in this way, yielding the sample $X_1, ..., X_n$. Sampling without replacement is sometimes called simple random sampling.

A sample drawn from a finite population without replacement does not satisfy the conditions of Definition 5.1. The random variables $X_1, ..., X_n$ are not mutually independent.

5.2 Sums of Random Variables from a Random Sample

Definition 5.4: Statistic / Sampling Distribution

Let $X_1, ..., X_n$ be a random sample of size n from a population and let $T(x_1, ..., x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, ..., X_n)$. Then the random variable or random vector $Y = T(X_1, ..., X_n)$ is called a statistic. The probability distribution of a statistic Y is called the sampling distribution of Y.

Definition 5.5: Sample Mean

Let $X_1, ..., X_n$ be a random sample. The sample mean is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \frac{X_1 + \cdots X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{5.3}$$

Definition 5.6: Sample Variance

Let $X_1, ..., X_n$ be a random sample. The sample variance is the statistic defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \tag{5.4}$$

where $\bar{X}$ is the sample mean.

Definition 5.7: Sample Standard Deviation

Let $X_1, ..., X_n$ be a random sample. The sample standard deviation is the statistic defined by $S = \sqrt{S^2}$ where S^2 is the sample variance.

Theorem 5.1

Let $x_1, ..., x_n$ be any numbers and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then

a. $min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$,

b. $(n-1)s^2 := \sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2$

Lemma 5.1

Let $X_1, ..., X_n$ be a random sample from a population and let $g(x)$ be a function such that $E[g(X_1)]$ and $Var[g(X_1)]$ exist. Then

$$E\left[\sum_{i=1}^{n} g(X_i)\right] = n\ E\big[g(X_1)\big] \tag{5.5}$$

and

$$Var\left[\sum_{i=1}^{n} g(X_i)\right] = n\ Var\big[g(X_1)\big] \tag{5.6}$$

Theorem 5.2

Let $X_1, ..., X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then

- **a.** $E[\bar{X}] = \mu$,
- **b.** $Var[\bar{X}] = \frac{\sigma^2}{n}$,
- **c.** $E[S^2] = \sigma^2$.

The relationships in Theorem 5.2 between a statistic and a population parameter, are examples of *unbiased statistics*. These will be further discussed in §7. The statistic $\bar{X}$ is an *unbiased estimator* of μ, and S^2 is an *unbiased estimator* of σ^2.

Proposition 5.2

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. If $f(y)$ is the PDF of $Y = (X_1 + \cdots + X_n)$, then $f_{\bar{X}}(x) = nf(nx)$ is the PDF of $\bar{X}$.

Theorem 5.3

Let $X_1, ..., X_n$ be a random sample from a population with MGF $M_X(t)$. Then the MGF of the sample mean is

$$M_{\bar{X}}(t) = \big[M_X(t/n)\big]^n \tag{5.7}$$

Theorem 5.4

If X and Y are independent continuous random variables with PDFs $f_X(x)$ and $f_Y(y)$, then the PDF of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(\omega) f_Y(z - \omega) d\omega \tag{5.8}$$

Theorem 5.5

Suppose $X_1, ..., X_n$ is a random sample from a PDF or PMF $f(x|\theta)$, where

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right) \tag{5.9}$$

is a member of an exponential family. Define statistics $T_1, ..., T_k$ by

$$T_i(X_1, ..., X_n) = \sum_{j=1}^{n} t_i(X_j), \qquad i = 1, ..., k \tag{5.10}$$

If the set $\{\big(w_1(\theta), w_2(\theta), ..., w_k(\theta)\big); \theta \in \Theta\}$ contains an open subset of $\mathbb{R}^k$, then the distribution of $(T_1, ..., T_k)$ is an exponential family of the form

$$f_T(u_1, ..., u_k|\theta) = H(u_1, ..., u_k)\big(c(\theta)\big)^n exp\Big(\sum_{i=1}^{k} w_i(\theta)u_i\Big) \tag{5.11}$$

5.3 Sampling from the Normal Distribution

5.3.1 Properties of the Sample Mean and Variance

Theorem 5.6

Let $X_1, ..., X_n$ be a random sample from a $n(\mu, \sigma^2)$ distribution, and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Then

- **a.** *$\bar{X}$ and S^2 are independent random variables.*
- **b.** *$\bar{X}$ has a $n(\mu, \sigma^2/n)$ distribution.*
- **c.** *$(n-1)S^2/\sigma^2$ has a chi squared distribution with $n-1$ degrees of freedom.*

Lemma 5.2: Facts about Chi Squared Random Variables

We use the notation χ^2_p to denote a chi squared random variable with p degrees of freedom.

- **a.** *If Z is a $n(0,1)$ random variable, then $Z^2 \sim \chi^2_1$; that is, the square of a standard normal random variable is a chi squared random variable.*
- **b.** *If $X_1, ..., X_n$ are independent and $X_i \sim \chi^2_{p_i}$, then $X_1 + \cdots + X_n \sim \chi^2_{p_1+\cdots+p_n}$; that is, independent chi squared variables add to a chi squared variable, and the degrees of freedom also add.*

Lemma 5.3

Let $X_j \sim n(\mu_j, \sigma_j^2), j = 1, ..., n$ be a collection of independent variables. For constants a_{ij} and b_{rj} $(j = 1, ..., n; i = 1, ..., k; r = 1, ..., m)$, where $k + m \leq n$. We define

$$U_i = \sum_{j=1}^{n} a_{ij}X_j, \qquad i = 1, ..., k \tag{5.12}$$

$$V_r = \sum_{j=1}^{n} b_{rj}X_j, \qquad r = 1, ..., m \tag{5.13}$$

- **a.** *The random variables U_i and V_r are independent if and only if $Cov(U_i, V_r) = 0$. Furthermore $Cov(U_i, V_r) = \sum_{j=1}^{n} a_{ij}b_{rj}\sigma_j^2$.*

b. *The random vectors $(U_1, ..., U_k)$ and $(V_1, ..., V_m)$ are independent if and only if U_i is independent of V_r for all pairs i, r $(i = 1, ..., k; r = 1, ..., m)$.*

This lemma shows that, if we start with independent normal random variables, covariance and independence are equivalent for linear functions of these random variables. Thus, we can check independence for normal variables by merely checking the covariance term, a much simpler calculation.

5.3.2 The Derived Distributions: Student's t and Snedecor's F

Definition 5.8: Student's t Distribution

Let $X_1, ..., X_n$ be a random sample from a $n(\mu, \sigma^2)$ distribution and S denote the sample standard deviation. The quantity $(X - \mu)/(S/\sqrt{n})$ has a Student's t distribution with $n - 1$ degrees of freedom. Equivalently, a random variable T has Student's t distribution with p degrees of freedom, and we write $T \sim t_p$ if it has PDF

$$f_T(t) = \frac{\Gamma\left(\frac{p-1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{(p\pi)^{1/2}} \frac{1}{(1 + t^2/p)^{(p+1)/2}}, \qquad -\infty < t < \infty \tag{5.14}$$

Proposition 5.3: Properties of t_p Distributed Variables

Let $T_p \sim t_p$. Then,

1. *$E[T_p] = 0$, if $p > 1$,*
2. *$Var[T_p] = \frac{p}{p-2}$, if $p > 2$.*
3. *$E[(T_p)^n]$ does not exist for $n \geq p$.*

Definition 5.9: Snedecor's F Distribution

Let $X_1, ..., X_n$ be a random sample from a $n(\mu_X, \sigma_X^2)$ population, and let $Y_1, ..., Y_m$ be a random sample from an independent $n(\mu_Y, \sigma_Y^2)$ population. The random variable $F = \left(S_X^2/\sigma_X^2\right)/\left(S_Y^2/\sigma_Y^2\right)$ has Snedecor's F distribution with $n - 1$ and $m - 1$ degrees of freedom. Equivalently, the random variable F has the F distribution with p and q degrees of freedom if it has PDF

$$f_F(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} \left(\frac{p}{q}\right)^{p/2} \frac{x^{p/2-1}}{[1 + (p/q)x]^{(p+q)/2}}, \qquad 0 < x < \infty \tag{5.15}$$

We denote the F distribution with p and q degrees of freedom by $F_{p,q}$.

Theorem 5.7: Properties of Snedecor's F Distribution

a. *If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$; that is, the reciprocal of an F random variable is again an F random variable.*

b. *If $X \sim t_q$, then $X^2 \sim F_{1,q}$.*

c. *If $X \sim F_{p,q}$, then $(p/q)X/(1 + (p/q)X) \sim \text{beta}(p/2, q/2)$*

5.4 Order Statistics

Definition 5.10: Order Statistics

The order statistics of a random sample $X_1, ..., X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, ..., X_{(n)}$. The order statistics are random variables that satisfy $X_{(1)} \leq \cdots \leq X_{(n)}$. In particular,

$$X_{(1)} = \min_{1 \leq i \leq n} X_i, \; X_{(2)} = \text{second smallest } X_i, \; \cdots, X_{(n)} = \max_{1 \leq i \leq n} X_i \tag{5.16}$$

Definition 5.11: Sample Range

Let $X_1, ..., X_n$ be a random sample and $X_{(1)}, ..., X_{(n)}$ denote the order statistics. Then, the sample range is the random variable $R = X_{(n)} - X_{(1)}$. It is the distance between the smallest and largest observations. It is a measure of the dispersion in the sample and should reflect the dispersion in the population.

Definition 5.12: Sample Median

The sample median, which we will denote by M, is a number such that approximately one-half of the observations are less than M and one-half are greater. In terms of the order statistics, M is defined by

$$M = \begin{cases} X_{((n+1)/2)}, & \text{if } n \text{ is odd} \\ (X_{(n/2)} + X_{(n/2+1)})/2 & \text{if } n \text{ is even} \end{cases} \tag{5.17}$$

Notation 5.1: $\{b\}$ Rounding

The notation $\{b\}$, when appearing in a subscript, is defined to be the number b rounded to the nearest integer in the usual way. More precisely, if i is an integer such that $i - 0.5 \leq b < i + 0.5$, then $\{b\} = i$.

Definition 5.13: (100p)th Sample Percentile

The $(100p)$th sample percentile is $X_{(\{np\})}$ if $\frac{1}{2n} < p < 0.5$ and $X_{(n+1-\{n(1-p)\})}$ if $0.5 < p < 1 - \frac{1}{2n}$.

Definition 5.14: Quartiles

We define the lower quartile as the 25th percentile and upper quartile as the 75th percentile. A measure of dispersion that is sometimes used is the interquartile range, the distance between the lower and upper quartiles.

Theorem 5.8

Let $X_1, .., X_n$ be a random sample from a discrete distribution with PMF $f_X(x_i) = p_i$, where $x_1 < x_2 < ...$

are the possible values of X in ascending order. Define

$$P_0 = 0 \tag{5.18}$$
$$P_1 = p_1 \tag{5.19}$$
$$P_2 = p_1 + p_2 \tag{5.20}$$
$$\vdots \tag{5.21}$$
$$P_i = \sum_{j=1}^{i} p_j \tag{5.22}$$
$$\vdots \tag{5.23}$$

Let $X_{(1)}, ..., X_{(n)}$ denote the order statistics from the sample. Then

$$P(X_{(j)} \leq x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k} \tag{5.24}$$

and

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}\right] \tag{5.25}$$

Theorem 5.9: Order Statistic Distribution

Let $X_{(1)}, ..., X_{(n)}$ denote the order statistics of a random sample, $X_1, ..., X_n$, from a continuous population with CDF $F_X(x)$ and PDF $f_X(x)$. Then the PDF of $X_{(j)}$ is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) \left[F_x(x)\right]^{j-1} \left[1 - F_X(x)\right]^{n-j}. \tag{5.26}$$

The CDF of $X_{(j)}$ is given by

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} \left[F_x(x)\right]^k \left[1 - F_X(x)\right]^{n-k} \tag{5.27}$$

Theorem 5.10

Let $X_{(1)}, ..., X_{(n)}$ denote the order statistics of a random sample $X_1, ..., X_n$, from a continuous population with CDF $F_X(x)$ and PDF $f_X(x)$. Then the joint PDF of $X_{(i)}$ and $X_{(j)}, 1 \leq i < j \leq n$ is

$$f_{X_{(i)},X_{(j)}}(u, v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v) \left[F_X(u)\right]^{i-1} \left[F_X(v) - F_X(u)\right]^{j-1-i} [1 - F_X(v)]^{n-j} \tag{5.28}$$

for $-\infty < u < v < \infty$.

Corollary 5.1

The joint PDF for all the order statistics is given by

$$f_{X_{(1)},..,X_{(n)}}(x_1, .., x_n) = \begin{cases} n! f_X(x_1) \cdots f_X(x_n) & -\infty < x_1 < \cdots < x_n < \infty \\ 0 & \text{otherwise} \end{cases} \tag{5.29}$$

5.5 Convergence Concepts

5.5.1 Convergence in Probability

Definition 5.15: Converges in Probability

A sequence of random variables, $X_1, X_2, ...$, converges in probability to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\big(|X_n - X| \geq \epsilon\big) = 0, \qquad \textit{or, equivalently,} \qquad \lim_{n\to\infty} P\big(|X_n - X| < \epsilon\big) = 1 \tag{5.30}$$

Theorem 5.11: Weak Law of Large Numbers

Let $X_1, X_2, ...$ be iid random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Then for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\big(|\bar{X}_n - \mu| < \epsilon\big) = 1, \tag{5.31}$$

that is, $\bar{X}_n$ converges in probability to μ.

Proof. We have that

$$P\big(|\bar{X}_n - \mu| \geq \epsilon\big) = P\big((\bar{X}_n - \mu)^2 \geq \epsilon^2\big) \leq \frac{E[(\bar{X}_n - \mu)^2]}{\epsilon^2}, \tag{5.32}$$

where the last inequality came from Chebyshev's Inequality. We recall that

$$E[(\bar{X}_n - \mu)^2] = \frac{\sigma^2}{n}. \tag{5.33}$$

Hence, we have that

$$0 \leq P\big(|\bar{X}_n - \mu| \geq \epsilon\big) \leq \frac{\sigma^2}{n\epsilon^2} \tag{5.34}$$

implying

$$\lim_{n\to\infty} 0 \leq \lim_{n\to\infty} P\big(|\bar{X}_n - \mu| \geq \epsilon\big) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0 \tag{5.35}$$

Hence, by the squeeze theorem, one has that

$$\lim_{n\to\infty} P\big(|\bar{X}_n - \mu| \geq \epsilon\big) = 0, \tag{5.36}$$

as desired. □

Theorem 5.12

Suppose that $X_1, X_2, ...$ converges in probability to a random variable X and that h is a continuous function. Then $h(X_1), h(X_2), ...$ converges in probability to $h(X)$.

5.5.2 Almost Sure Convergence

A type of convergence that is stronger than convergence in probability is almost sure convergence (sometimes known as *convergence with probability* 1). This type of convergence is similar to pointwise convergence of a sequence of functions, except that the convergence need not occur on a set with probability 0.

Definition 5.16: Converges Almost Surely

A sequence of random variables, $X_1, X_2, ...$, converges almost surely to a random variable X if, for every $\epsilon > 0$,

$$P\big(\lim_{n\to\infty} |X_n - X| < \epsilon\big) = 1 \tag{5.37}$$

Theorem 5.13: Strong Law of Large Numbers

Let $X_1, X_2, ...$ be iid random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$, and define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Then, for every $\epsilon > 0$,

$$P\big(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon\big) = 1; \tag{5.38}$$

that is, $\bar{X}_n$ converges almost surely to μ.

For both the Weak and Strong Laws of Large Numbers, we had the assumption of a finite variance. However, these laws hold without this assumption on a finite variance. The only moment condition that is needed is $E[|X_i|] < \infty$.

5.5.3 Convergence in Distribution

Definition 5.17: Converges in Distribution

A sequence of random variables $X_1, X_2, ...$ converges in distribution to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \tag{5.39}$$

at all points x where $F_X(x)$ is continuous.

Theorem 5.14

If the sequence of random variables $X_1, X_2, ...$ converges in probability to a random variable X, the sequence also converges in distribution to X.

Theorem 5.15

The sequence of random variables $X_1, X_2, ...$ converges in probability to a constant μ if and only if the sequence also converges in distribution to μ. That is, the statement

$$P(|X_n - \mu| > \epsilon) \to 0 \text{ for every } \epsilon > 0 \tag{5.40}$$

is equivalent to

$$P(X_n \le x) \to \begin{cases} 0 & \text{if } x < \mu \\ 1 & \text{if } x > \mu \end{cases} \tag{5.41}$$

Theorem 5.16: Central Limit Theorem

Let $X_1, X_2, ...$ be a sequence of iid random variables whose MGFs exist in a neighbourhood of 0 (that is, $M_{X_i}(t)$ exists for $|t| < h$, for some positive h). Let $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 > 0$. (Both μ and σ^2 are

finite since the MGF exists.) Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. *Let* $G_n(x)$ *denote the CDF of* $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. *Then, for any* $x, -\infty < x < \infty$,

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy; \tag{5.42}$$

that is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ *has a limiting standard normal distribution.*

Theorem 5.17: Stronger Form of the Central Limit Theorem

Let $X_1, X_2, \ldots$ *be a sequence of iid random variables with* $E[X_i] = \mu$ *and* $0 < Var[X_i] = \sigma^2 < \infty$. *Define* $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. *Let* $G_n(x)$ *denote the CDF of* $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. *Then, for any* $x, -\infty < x < \infty$,

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \tag{5.43}$$

that is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ *has a limiting standard normal distribution.*

Theorem 5.18: Slutsky's Theorem

If $X_n \to X$ *in distribution and* $Y_n \to a$ *in probability (where* a *is a constant), then*

- **a.** $Y_n X_n \to aX$ *in distribution.*
- **b.** $X_n + Y_n \to X + a$ *in distribution.*

5.5.4 The Delta Method

Definition 5.18: Taylor Polynomial

If a function $g(x)$ *has derivatives of order* r, *that is,* $g^{(r)}(x) = \frac{d^r}{dx^r} g(x)$ *exists, then for any constant* a, *the Taylor polynomial of order* r *about* a *is*

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i \tag{5.44}$$

Theorem 5.19: Taylor's Theorem

If $g^{(r)}(a) = \frac{d^r}{dx^r} g(x)|_{x=a}$ *exists, then*

$$\lim_{x\to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0 \tag{5.45}$$

In essence, the theorem says that the *remainder* from the approximation, $g(x) = T_r(x)$ always tends to 0 faster than the highest-order explicit term.

Proposition 5.4

Let $g : \mathbb{R} \to \mathbb{R}$ *be a function whose (r+1)th derivative exists. Let* T_r *denote the Taylor polynomial of degree*

r. Then,

$$g(x) - T_r(x) = \int_a^x \frac{g^{(r+1)}(t)}{r!}(x-t)^r dt \tag{5.46}$$

Statistical Approximation via Taylor Series

Let $T_1, ..., T_k$ be random variables with means $\theta_1, ..., \theta_k$, and define $\mathbf{T} = (T_1, ..., T_k)$ and $\theta = (\theta_1, ..., \theta_k)$. Suppose that there is a differentiable function $g(\mathbf{T})$ (an estimator of some parameter) for which we want an approximate estimate of variance. Define

$$g_i'(\theta) = \frac{\partial}{\partial t_i} g(\mathbf{t})\Big|_{t_1=\theta_1,...,t_k=\theta_k} \tag{5.47}$$

The first-order Taylor series expansion of g about θ is

$$g(t) = g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(t_i - \theta_i) + \textit{Remainder} \tag{5.48}$$

For our statistical approximation we forget about the remainder and write

$$g(t) \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(t_i - \theta_i) \tag{5.49}$$

Theorem 5.20: Delta Method

Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to n(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and non-zero. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \to n(0, \sigma^2[g'(\theta)]^2) \textit{ in distribution} \tag{5.50}$$

Theorem 5.21: Second-Order Delta Method

Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to n(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then

$$n[g(Y_n) - g(\theta)] \to \sigma^2 \frac{g''(\theta)}{2} \chi_1^2 \textit{ in distribution} \tag{5.51}$$

Theorem 5.22: Multivariate Delta Method

Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be a random sample with $E[X_{ij}] = \mu_i$ and $Cov(X_{ik}, X_{jk}) = \sigma_{ij}$. For a given function g with continuous first partial derivatives and a specific value of $\mu = (\mu_1, ..., \mu_p)$ for which $\tau^2 = \sum\sum \sigma_{ij} \frac{\partial g}{\partial \mu_i} \frac{\partial g}{\partial \mu_j} > 0$,

$$\sqrt{n}\left[g(\bar{X}_1, ..., \bar{X}_s) - g(\mu_1, ..., \mu_p)\right] \to n(0, \tau^2) \textit{ in distribution} \tag{5.52}$$

5.6 Generating a Random Sample

Example 5.1: Probability Integral Transform

If Y is a continuous random variable with CDF F_Y, then Theorem 2.4 implies that the random variable $F_Y^{-1}(U)$, where $U \sim$ uniform$(0,1)$ has distribution F_Y. For instance, consider $Y \sim$ exponential(λ), then

$$F_Y^{-1}(U) = -\lambda \log(1-U) \tag{5.53}$$

is an exponential(λ) random variable. Hence, if we generate $U_1, ..., U_n$ as iid random variables, $Y_i = -\lambda \log(1-U_i)$ $(1 \leq i \leq n)$ are iid exponential(λ) random variables. The relationship between exponential and other distributions allows the quick generation of many random variables.

Example 5.2: Box-Muller Algorithm

Let U_1 and U_2 be two generated independent uniform$(0,1)$ random variables, and set

$$R = \sqrt{-2log(U_1)} \text{ and } \theta = 2\pi U_2. \tag{5.54}$$

Then

$$X = R\cos(\theta) \text{ and } Y = R\sin(\theta) \tag{5.55}$$

are independent normal$(0,1)$ random variables. Thus, although there is no quick transformation for generating a single $n(0,1)$ random variable, there is such a method for generating two variables.

5.6.1 The Accept / Reject Algorithm

Theorem 5.23: Accept / Reject Algorithm

Let $Y \sim f_Y(y)$ and $V \sim f_V(v)$, where f_Y and f_V have common support with

$$M = \sup_y \frac{f_Y(y)}{f_V(y)} < \infty \tag{5.56}$$

To generate a random variable $Y \sim f_Y$:

- **a.** *Generate $U \sim$ uniform$(0,1)$, $V \sim f_V$, independent.*
- **b.** *If $U < \frac{1}{M} f_Y(V)/f_V(V)$, set $Y = V$; otherwise, return to step (a).*

Metropolis Algorithm

Let $Y \sim f_Y(y)$ and $V \sim f_V(v)$, where f_Y and f_V have common support. To generate $Y \sim f_Y$:

0. Generate $V \sim f_V$. Set $Z_0 = V$.

For $i = 1, 2, ...$:

1. Generate $U_i \sim$ uniform$(0,1), V_i \sim f_V$, and calculate

$$\rho_i = min\left\{ \frac{f_Y(V_i)}{f_V(V_i)} \frac{f_V(Z_{i-1})}{f_Y(Z_{i-1})}, 1 \right\} \tag{5.57}$$

2. *Set*

$$Z_i = \begin{cases} V_i & \text{if } U_i \leq \rho_i \\ Z_{i-1} & \text{if } U_i > \rho_i \end{cases} \tag{5.58}$$

Then, as $i \to \infty$, Z_i converges to Y in distribution.

Although the algorithm does not require a finite M, it does not produce a random variable with exactly the density f_Y, but rather a convergent sequence. In practice, after the algorithm runs for a while (i gets big), the Zs that are produced behave very much like variable from f_Y.

5.7 Miscellanea

Definition 5.19: Markov Chain

A sequence of random variables $X_1, X_2, ...$ is a Markov Chain if

$$P(X_{k+1} \in A | X_1, ..., X_k) = P(X_{k+1} \in A | X_k), \tag{5.59}$$

that is, the distribution of the present random variables depends, at most, on the immediate past random variable.

Theorem 5.24: Ergodic Theorem

If a Markov chain $X_1, X_2, ...$ satisfies some regularity conditions, then

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i) \to E[h(X)] \text{ as } n \to \infty, \tag{5.60}$$

provided that the expectation exists.

Methods that are collectively known as Markov Chain Monte Carlo (MCMC) methods are used in the generation of random variables and have proved extremely useful for doing complicated calculations, most notably, calculations involving integrations and maximizations.

Chapter 6

Principles of Data Reduction

Any statistic, $T(\mathbf{X})$ defines a form of data reduction or data summary. An experimenter who uses only the observed value of the statistic, $T(\mathbf{x})$, rather than the entire observed sample, $\mathbf{x}$, will treat as equal two samples, $\mathbf{x}$ and $\mathbf{y}$, that satisfy $T(\mathbf{x}) = T(\mathbf{y})$ even though the actual sample values may be different in some ways.

Data reduction in terms of a particular statistic can be thought of as a partition of the sample space $\mathcal{X}$. Let $\mathcal{T} = \{t : t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\mathbf{x})$. Then $T(\mathbf{x})$ partitions the sample space into sets $A_t, t \in \mathcal{T}$, defined by $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$. The statistic summarizes the data in that, rather than reporting the entire sample $\mathbf{x}$, it reports only that $T(\mathbf{x}) = t$ or, equivalently, $\mathbf{x} \in A_t$.

We'll study three principles of data reduction. We are interested in methods of data reduction that do not discard important information about the unknown parameter θ and methods that successfully discard information that is irrelevant as far as gaining knowledge about θ is concerned.

1. The Sufficiency Principle promotes a method of data reduction that does not discard information about θ while achieving some summarization of the data.

2. The Likelihood Principle describes a function of the parameter, determined by the observed sample, that contains all the information about θ that is available from the sample.

3. The Equivariance Principle prescribes yet another method of data reduction that still preserves some important features of the model.

6.1 The Sufficiency Principle

Informally, a *sufficient statistic* for a parameter θ is a statistic that, in a certain sense, captures all the information about θ contained in the sample.

Sufficiency Principle

If $T(\mathbf{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value $T(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

6.1.1 Sufficient Statistics

Definition 6.1: Sufficient Statistic

A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ.

Understanding Sufficient Statistics

Let t be a possible value of $T(\mathbf{X})$, that is, a value such that $P_\theta(T(\mathbf{X}) = t) > 0$. We consider the conditional probability $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t)$. If $\mathbf{x}$ is a sample point such that $T(\mathbf{x}) \neq t$, then it's clear that $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t) = 0$. We are therefore interested in $P\big(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})\big)$. By definition, if $T(\mathbf{X})$ is a sufficient statistic, this conditional probability is the same for all values of θ so we omitted the subscript.

Theorem 6.1

If $p(\mathbf{x}|\theta)$ is the joint PDF or PMF of $\mathbf{X}$ and $q(t|\theta)$ is the PDF or PMF of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{x}$ in the sample space, the ratio $p(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$ is constant as a function of θ.

Theorem 6.2: Factorization Theorem

Let $f(\mathbf{x}|\theta)$ denote the joint PDF or PMF of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(\mathbf{x})$ such that, for all sample points $\mathbf{x}$ and all parameter points θ,

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}) \tag{6.1}$$

Theorem 6.3

Let $X_1, ..., X_n$ be iid observations from a PDF or PMF $f(x|\theta)$ that belongs to an exponential family given by

$$f(x|\theta) = h(x)c(\theta)exp\bigg(\sum_{i=1}^{k} w_i(\theta)t_i(x)\bigg), \tag{6.2}$$

where $\theta = (\theta_1, \theta_2, ..., \theta_d), d \leq k$. Then

$$T(\mathbf{X}) = \bigg(\sum_{j=1}^{n} t_1(X_j), ..., \sum_{j=1}^{n} t_k(X_j)\bigg) \tag{6.3}$$

is a sufficient statistic for θ.

6.1.2 Minimal Sufficient Statistics

Definition 6.2: Minimal Sufficient Statistic

A sufficient statistic $T(\mathbf{X})$ is called a minimal sufficient statistic if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{X})$ is a function of $T'(\mathbf{x})$.[a]

[a] To say that $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ simply means that if $T'(\mathbf{x}) = T'(\mathbf{y})$, then $T(\mathbf{x}) = T(\mathbf{y})$. Alternatively, we say that $T(x)$ is a function of $T'(x)$ if there exists a function f such that $T(x) = f(T'(x))$.

Theorem 6.4: Testing Minimal Sufficiency

Let $f(\mathbf{x}|\theta)$ be the PMF or PDF of a sample $\mathbf{X}$. Suppose that there exists a function $T(\mathbf{x})$ such that, for every two samples points $\mathbf{x}$ and $\mathbf{y}$, the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is a minimal sufficient statistic for θ.

6.1.3 Sufficient, Ancillary, and Complete Statistics

Definition 6.3: Ancillary Statistic

A statistic $S(\mathbf{X})$ whose distribution does not depend on the parameter θ is called an ancillary statistic (relative to the parameter θ).

Definition 6.4: Complete Statistic

Let $f(t|\theta)$ be a family of PDFs or PMFs for a statistic $T(\mathbf{X})$. The family of probability distributions is called complete if $E_\theta[g(T)] = 0$ for all θ implies $P_\theta(g(T) = 0) = 1$ for all θ. Equivalently, $T(\mathbf{X})$ is called a complete statistic.

Theorem 6.5: Basu's Theorem

If $T(\mathbf{X})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic.

Theorem 6.6: Complete Statistics in the Exponential Family

Let $X_1, ..., X_n$ be iid observations from an exponential family with PDF or PMF of the form

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^{k} w(\theta_j)t)j(x)\right), \tag{6.4}$$

where $\theta = (\theta_1, ..., \theta_k)$. Then the statistic

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_i), ..., \sum_{i=1}^{n} t_k(X_i)\right) \tag{6.5}$$

is complete as long as the parameter space Θ contains an open set in $\mathbb{R}^k$.

Theorem 6.7

If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

6.2 The Likelihood Principle

6.2.1 The Likelihood Function

Definition 6.5: Likelihood Function

Let $f(\mathbf{x}|\theta)$ denote the joint PDF or PMF of the samples $\mathbf{X} = (X_1, ..., X_n)$. Then, given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) \tag{6.6}$$

is called the likelihood function.

LIKELIHOOD PRINCIPLE

If $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $L(\theta|\mathbf{x})$ is proportional to $L(\theta|\mathbf{y})$, that is, there exists a constant $C(\mathbf{x}, \mathbf{y})$ such that

$$L(\theta, \mathbf{x}) = C(\mathbf{x}, \mathbf{y})L(\theta|\mathbf{y}) \text{ for all } \theta, \tag{6.7}$$

then the conclusions drawn from $\mathbf{x}$ and $\mathbf{y}$ should be identical.

In the special case of $C(\mathbf{x}, \mathbf{y}) = 1$, the Likelihood Principle states that if two sample points result in the same likelihood function, then they contain the same information about θ.

6.2.2 The Formal Likelihood Principle

Definition 6.6: Experiment and Evidence

Formally, we define an experiment as a triple $(\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$, where $\mathbf{X}$ is a random vector with PMF $f(\mathbf{x}|\theta)$ for some θ in the parameter space Θ. An experimenter, knowing what experiment E was performed and having observed a particular sample $\mathbf{X} = \mathbf{x}$, will make some inference or draw some conclusion about θ. This conclusion we denote by $Ev(E, \mathbf{x})$, which stands for the evidence about θ arising from E and $\mathbf{x}$.

FORMAL SUFFICIENCY PRINCIPLE

Consider experiment $E = \big(\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\}\big)$ and suppose $T(\mathbf{X})$ is a sufficient statistic for θ. If $\mathbf{x}$ and $\mathbf{y}$ are sample points satisfying $T(\mathbf{x}) = T(\mathbf{y})$, then $Ev(E, \mathbf{x}) = Ev(E, \mathbf{y})$.

The Formal Sufficiency Principle goes a bit further than the Sufficiency Principle stated in the previous section. Here, we are agreeing to equate evidence if the sufficient statistics match.

CONDITIONALITY PRINCIPLE

Suppose that $E_1 = \big(\mathbf{X}_1, \theta, \{f_1(\mathbf{x}_1|\theta)\}\big)$ and $E_2 = \big(\mathbf{X}_2, \theta, \{f_2(\mathbf{x}_2|\theta)\}\big)$ are two experiments, where only the unknown parameter θ need be common between the two experiments. Consider the mixed experiment in which the random variable J is observed, where $P(J = 1) = P(J = 2) = \frac{1}{2}$ (independent of $\theta, \mathbf{X}_1$ or $\mathbf{X}_2$), and then experiment E_J is performed. Formally, the experiment performed is $E^* = (\mathbf{X}^*, \theta, \{f^*(\mathbf{x}^*|\theta)\})$, where $\mathbf{X}^* = (j, X_j)$ and

$$f^*(\mathbf{x}^*|\theta) = f^*((j, \mathbf{x}_j|\theta) = \frac{1}{2} f_j(\mathbf{x}_j|\theta). \tag{6.8}$$

Then

$$Ev(E^*, (j, x_j)\big) = Ev(E_j, \mathbf{x}_j). \tag{6.9}$$

The Conditionality Principle states quite simply that if one of two experiments is randomly chosen and the chosen experiment is done, yielding data $\mathbf{x}$, the information about θ *depends only on the experiment performed*. It is the same information as would have been obtained if it were decided (nonrandomly) to do that experiment from the beginning, and data $\mathbf{x}$ had been observed. The fact that this experiment was performed, rather than some other, has not increased, decreased, or changed knowledge of θ.

FORMAL LIKELIHOOD PRINCIPLE

Suppose that we have two experiments, $E_1 = (\mathbf{X}_1, \theta, \{f_1(\mathbf{x}_1|\theta)\})$ and $E_2 = (\mathbf{X}_2, \theta, \{f_2(\mathbf{x}_2|\theta)\})$, where the unknown parameter θ is the same in both experiments. Suppose $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ are sample points from E_1 and E_2, respectively, such that

$$L(\theta|\mathbf{x}_2^*) = CL(\theta|\mathbf{x}_1^*) \tag{6.10}$$

for all θ and for some constant C that may depend on $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ but not θ. Then

$$Ev(E_1, \mathbf{x}_1^*) = Ev(E_2, \mathbf{x}_2^*) \tag{6.11}$$

LIKELIHOOD PRINCIPLE COROLLARY

If $E = \left(\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\}\right)$ is an experiment then $Ev(E, \mathbf{x})$ should depend on E and $\mathbf{x}$ only through $L(\theta|\mathbf{x})$.

Theorem 6.8: Birnbaum's Theorem

The formal likelihood principle follows from the formal sufficiency principle and the conditionality principle. The converse is also true.

Since many intuitively appealing inference procedures do violate the Likelihood Principle, it is not universally accepted by all statisticians. Yet it is mathematically appealing and does suggest a useful data reduction technique.

6.3 The Equivariance Principle

Definition 6.7: Measurement Equivariance

Measurement Equivariance prescribes that the inference made should not depend on the measurement scale that is used.

Definition 6.8: Formal Invariance

Formal Invariance states that if two inference problems have the same formal structure in terms of the mathematical model used, then the same inference procedure should be used in both problems. The elements of the model that must be the same are: Θ, the parameter space; $\{f(\mathbf{x}|\theta) : \theta \in \Theta\}$, the set of PDFs or PMFs for the sample; and the set of allowable inferences and consequences of wrong inferences. An inference is simply a choice of an element of Θ as an estimate or guess at the true value of θ.

EQUIVARIANCE PRINCIPLE

If $\mathbf{Y} = g(\mathbf{X})$ is a change of measurement scale such that the model for $\mathbf{Y}$ has the same formal structure as the model for $\mathbf{X}$, then an inference procedure should be both measurement equivariant and formally equivariant.

Definition 6.9: Group of Transformations

A set of functions $\{g : \mathcal{X} \to \mathcal{X}; g \in \mathcal{G}\}$ from the "sample space" $\mathcal{X}$ onto $\mathcal{X}$ is called a group of transformations of $\mathcal{X}$ if

(i) (Inverse) For every $g \in \mathcal{G}$ there is a $g' \in \mathcal{G}$ such that $g'(g(\mathbf{x})) = \mathbf{x}$ for all $\mathbf{x} \in \mathcal{X}$.

(ii) (Composition) For every $g \in \mathcal{G}$ and $g' \in \mathcal{G}$ there exists $g'' \in \mathcal{G}$ such that $g'(g(\mathbf{x})) = g''(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$.

Sometimes the third requirement,

(iii) (Identity) The identity, $e(\mathbf{x})$, defined by $e(\mathbf{x}) = \mathbf{x}$ is an element of $\mathcal{G}$,

is stated as part of the definition of a group. However, (iii) is a consequence of (i) and (ii).

Definition 6.10: Invariance Under Group Transformation

Let $\mathcal{F} = \{f(\mathbf{x}|\theta) : \theta \in \Theta\}$ be a set of PDFs or PMFs for $\mathbf{X}$, and let $\mathcal{G}$ be a group of transformations of the sample space $\mathcal{X}$. Then $\mathcal{F}$ is invariant under the group $\mathcal{G}$ if for every $\theta \in \Theta$ and $g \in \mathcal{G}$ there exists a unique $\theta' \in \Theta$ such that $Y = G(\mathbf{X})$ has the distribution $f(\mathbf{y}|\theta')$ if $\mathbf{X}$ has the distribution $f(\mathbf{x}|\theta)$.

6.4 Miscellanea

Definition 6.11: First-Order Ancillary

A statistic $V(\mathbf{X})$ is called first-order ancillary if $E_\theta[V(\mathbf{X})]$ is independent of θ.

Theorem 6.9

Let T be a statistic with $Var[T] < \infty$. A necessary and sufficient condition for T to be complete is that every bounded first-order ancillary V is uncorrelated (for all θ) with every bounded real-valued function of T.

Definition 6.12: Necessary Statistic

A statistic is said to be necessary if it can be written as a function of every sufficient statistic.

Theorem 6.10

A statistic is a minimal sufficient statistic if and only if it is a necessary and sufficient statistic.

Theorem 6.11: Minimal Sufficient Statistics

Suppose that the family of densities $\{f_0(\mathbf{x}), ..., f_k(\mathbf{x})\}$ all have common support. Then

a. *The statistic*

$$T(\mathbf{X}) = \left(\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})}, \frac{f_2(\mathbf{X})}{f_0(\mathbf{X})}, ..., \frac{f_k(\mathbf{X})}{f_0(\mathbf{X})}\right) \tag{6.12}$$

is minimal sufficient for the family $\{f_0(\mathbf{x}), ..., f_k(\mathbf{x})\}$.

b. *If $\mathcal{F}$ is a family of densities with common support, and*

(i) $f_i(\mathbf{x}) \in \mathcal{F}$, $i = 0, 1, ..., k$,

(ii) $T(\mathbf{x})$ *is sufficient for* $\mathcal{F}$,

then $T(\mathbf{x})$ *is minimal sufficient for* $\mathcal{F}$.

Chapter 7

Point Estimation

Definition 7.1: Point Estimator

A point estimator is any function $W(X_1, ..., X_n)$ of a sample; that is, any statistic is a point estimator.

Definition 7.2: Estimator and Estimate

An estimator is a function of the sample, while an estimate is the realized value of an estimator (i.e a scalar) that is obtained when a sample is actually taken. Notationally, when a sample is taken, an estimator is a function of the random variables $X_1, .., X_n$ while an estimate is a function of the realized values $x_1, ..., x_n$.

7.1 Methods of Finding Estimators

7.1.1 Method of Moments

Let $X_1, ..., X_n$ be a sample from a population with PDF or PMF $f(x|\theta_1, ..., \theta_k)$. Method of moments estimators are found by equating the first k sample moments to the corresponding k population moments, and solving the resulting system of simultaneous equations. More precisely, we define

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^1, \qquad \mu_1' = E[X^1], \tag{7.1}$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \qquad \mu_2' = E[X^2], \tag{7.2}$$

$$\vdots \tag{7.3}$$

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad \mu_k' = E[X^k] \tag{7.4}$$

The method of moments estimator $(\bar{\theta}_1, ..., \bar{\theta}_k)$ of $(\theta_1, ..., \theta_k)$ is obtained by solving the following system of equations for $(\theta_1, ..., \theta_k)$ in terms of $(m_1, ..., m_k)$:

$$m_1 = \mu_1'(\theta_1, ..., \theta_k), \tag{7.5}$$

$$m_2 = \mu_2'(\theta_1, ..., \theta_k), \tag{7.6}$$

$$\vdots \tag{7.7}$$

$$m_k = \mu_k'(\theta_1, ..., \theta_k) \tag{7.8}$$

7.1.2 Maximum Likelihood Estimators

Definition 7.3: Maximum Likelihood Estimator (MLE)

For each sample point $\mathbf{x}$*, let* $\hat{\theta}(\mathbf{x})$ *be a parameter value at which* $L(\theta|\mathbf{x})$ *attains its maximum as a function of* θ*, with* $\mathbf{x}$ *held fixed. A maximum likelihood estimator (MLE) of the parameter* θ *based on a sample* $\mathbf{X}$ *is* $\hat{\theta}(\mathbf{X})$*.*

We first remind ourselves of a key result in multivariable calculus.

Definition 7.4: Critical Points

Let $f : U \to \mathbb{R}$ *be a continuous function where* $U \subset \mathbb{R}^n$*. Let* ∇ *denote the gradient operator on Euclidean space. Then,* $\mathbf{x}$ *is said to be a critical point of* f *if*

$$\nabla f(\mathbf{x}) = \mathbf{0} \tag{7.9}$$

or if $\nabla f(\mathbf{x})$ *does not exist.*[a]

[a] Some authors exclude the non-existence scenario from this definition.

Theorem 7.1: Extrema Points on Compact Domains

Let $U \subset \mathbb{R}^n$ *be a compact (closed and bounded) set. If* $f : U \to \mathbb{R}$ *is continuous everywhere, then its extrema (i.e global maximum and global minimum) occur either on the boundary of* U *or at the interior critical points of* f*.*[a]

[a] The critical points may also occur on the boundary, but we have already included this consideration with the general statement of "*on the boundary of* U".

Definition 7.5: Possible MLE Candidates

Let $L(\theta|\mathbf{x})$ *be the likelihood function. If the likelihood function is differentiable in* θ_i*, then possible candidates for MLE are the values* $(\theta_1, ..., \theta_k)$ *that solve*

$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}), \qquad i = 1, ..., k. \tag{7.10}$$

The solutions are only possible candidates for MLE *since first derivative being* 0 *is only a necessary condition for a maximum, not a sufficient condition. The extrema may occur on the boundary, which would have to be checked seperately.*

Definition 7.6: Induced Likelihood Function

Let Θ *denote the parameter space for* θ*. Let* $\tau : \Theta \to \Sigma$*, where* Σ *denotes a new parameter space for* $\eta = \tau(\theta)$*. Let* L *be the likelihood function for* θ*. Then, the induced likelihood function for* η *is defined as*

$$L^*(\eta|\mathbf{x}) = \sup_{\{\theta:\tau(\theta)=\eta\}} L(\theta|\mathbf{x}) \tag{7.11}$$

The value $\hat{\eta}$ *that maximizes* $L^*(\eta|\mathbf{x})$ *will be called the MLE of* $\eta = \tau(\theta)$*.*

Theorem 7.2: Invariance Property of MLEs

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

7.1.3 Bayes Estimators

Definition 7.7: Prior and Posterior Distribution

In the Bayesian approach θ is considered to be a quantity whose variation can be described by a probability distribution (called the prior distribution). This is a subjective description, based on the experimenter's belief, and is formulated before the data are seen. A sample is then taken from a population indexed by θ and the prior distribution is updated with this sample information. The updated prior is called the posterior distribution. Let $\pi(\theta)$ denote the prior distribution and the sampling distribution by $f(\mathbf{x}|\theta)$. Then the posterior distribution, which is the conditional distribution of θ given the sample, $\mathbf{x}$, is given by

$$\pi(\theta, \mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}, \tag{7.12}$$

where $m(\mathbf{x})$ is the marginal distribution of $\mathbf{X}$:

$$m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)\ d\theta \tag{7.13}$$

Definition 7.8: Conjugate Family

Let $\mathcal{F}$ denote the class of PDFs or PMFs $f(x|\theta)$ (indexed by θ). A class Π of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π, and all $x \in \mathcal{X}$.

7.1.4 The Expectation-Maximization (EM) Algorithm

7.2 Methods of Evaluating Estimators

The methods discussed in the previous section have outlined reasonable techniques for finding point estimators of parameters. A difficulty that arises, however, is that since we can usually apply more than one of these methods in a particular situation, we are often faced with the task of choosing between estimators.

The general topic of evaluating statistical procedures is part of the branch of statistics known as decision theory.

7.2.1 Mean Squared Error

Definition 7.9: Mean Squared Error

The mean squared error (MSE) of an estimator W of a parameter θ is the function of θ defined by $E_\theta[(W-\theta)^2]$

Definition 7.10: Bias

The bias of a point estimator W of a parameter θ is the difference between the expected value of W and θ; that is, $Bias_\theta(W) = E_\theta[W] - \theta$. An estimator whose bias is identically (in θ) equal to 0 is called unbiased and satisfies $E_\theta[W] = \theta$ for all θ.

Theorem 7.3: Bias-Variance Decomposition

Let W be an estimator for θ. Then, its mean squared error can be decomposed into bias and variance components:

$$E_\theta[(W-\theta)^2] = Var_\theta[W] + (Bias_\theta[W])^2 \tag{7.14}$$

7.2.2 Best Unbiased Estimators

The reason that there is no one "*best MSE*" estimator is that the class of all estimators is too large a class. One way to make the problem of finding a "*best*" estimator tractable is to limit the class of estimators. A popular way of restricting the class of estimators, the one we consider in this section, is to consider only unbiased estimators.

Definition 7.11: Best Unbiased Estimator

An estimator W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies $E_\theta[W^*] = \tau(\theta)$ for all θ and, for any other estimator W with $E_\theta[W] = \tau(\theta)$, we have $Var_\theta[W^*] \leq Var_\theta[W]$ for all θ. W^* is also called a uniform minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$.

In essence, the class of best unbiased estimators $\{W\}$ for $\tau(\theta)$ are those that minimize the mean squared error $E_\theta[(W-\theta)^2]$ given that $E_\theta[W] = \tau(\theta)$.

Theorem 7.4: Cramer-Rao Inequality

Let $X_1, ..., X_n$ be a sample (not necessarily iid X_i) with PDF $f(\mathbf{x}|\theta)$, and let $W(\mathbf{X}) = W(X_1, ..., X_n)$ be an estimator satisfying

$$\frac{d}{d\theta}E_\theta[W(\mathbf{X})] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}[W(\mathbf{x})f(\mathbf{x}|\theta)]d\mathbf{x} \tag{7.15}$$

and

$$Var_\theta[W(\mathbf{X})] < \infty. \tag{7.16}$$

Then

$$Var_\theta[W(\mathbf{X})] \geq \frac{\frac{d}{d\theta}E_\theta[W(\mathbf{X})]}{E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right]} \tag{7.17}$$

Corollary 7.1: Cramer-Rao Inequality: iid Case

If the assumptions of Theorem 7.4 are satisfied and, additionally, if $X_1, ..., X_n$ are iid with PDF $f(x|\theta)$, then

$$Var_\theta[W(\mathbf{X})] \geq \frac{\left(\frac{d}{d\theta}E_\theta[W(\mathbf{X})]\right)^2}{nE_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]} \tag{7.18}$$

Definition 7.12: Information Number / Fisher Information

Consider a sample $X_1, ..., X_n$ with joint PDF $f(\mathbf{x}|\theta)$. The quantity

$$E_\theta\left[\left(\frac{\partial}{\partial\theta} log\ f(\mathbf{X}|\theta)\right)^2\right] \tag{7.19}$$

is called the information number, or Fisher information of the sample.

This terminology reflects the fact that the information number gives a bound on the variance of the best unbiased estimator of θ. As the information number gets bigger and we have more information about θ, we have a smaller bound on the variance of the best unbiased estimator.

Lemma 7.1

If $f(x|\theta)$ satisfies

$$\frac{d}{d\theta}E_\theta\left[\frac{\partial}{\partial\theta} log f(X|\theta)\right] = \int \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta} log\ f(x|\theta)\right) f(x|\theta)\right] dx \tag{7.20}$$

(true for an exponential family), then

$$E_\theta\left[\left(\frac{\partial}{\partial\theta} log\ f(X|\theta)\right)^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2} log\ f(X|\theta)\right] \tag{7.21}$$

A shortcoming of this approach to finding best unbiased estimators is that, even if the Cramer-Rao Theorem is applicable, there is no guarantee that the bound is sharp. That is to say, the value of the Cramer-Rao lower bound may be strictly smaller than the variance of any unbiased estimator.

Corollary 7.2: Lower Bound Attainment

Let $X_1, ..., X_n$ be iid with PDF $f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the Cramer-Rao Theorem. Let $L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta)$ denote the likelihood function. If $W(\mathbf{X}) = W(X_1, ..., X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramer-Rao Lower Bound if and only if

$$a(\theta)\big[W(\mathbf{x}) - \tau(\theta)\big] = \frac{\partial}{\partial\theta} log\ L(\theta|\mathbf{x}) \tag{7.22}$$

for some function $a(\theta)$.

7.2.3 Sufficiency and Unbiasedness

Theorem 7.5: Rao-Blackwell

Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient statistic for θ. Define $\phi(T) = E[W|T]$. Then $E_\theta[\phi(T)] = \tau(\theta)$ and $Var_\theta[\phi(T)] \leq Var_\theta[W]$ for all θ; that is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.

Theorem 7.6

If W is a best unbiased estimator of $\tau(\theta)$, then W is unique.

Definition 7.13: Unbiased Estimator of 0

An estimator U is said to be an unbiased estimator of 0 if it satisfies $E_\theta[U] = 0$ for all θ.[a]

[a] With the terminology introduced thus far, this definition appears redundant and you would be absolutely right. However, I refer to such an unbiased estimator of 0 a few times and wanted to make clear that it is what you think it is.

Theorem 7.7

If $E_\theta[W] = \tau(\theta)$, W is the best unbiased estimator of $\tau(\theta)$ if and only if W is uncorrelated with all unbiased estimators of 0.

Theorem 7.8

Let T be a complete sufficient statistic for a parameter θ, and let $\phi(T)$ be any estimator based only on T. Then $\phi(T)$ is the unique best unbiased estimator of its expected value.

7.2.4 Loss Function Optimality

Definition 7.14: Decision Theory

Our evaluations of point estimators have been based on their mean squared error performance. Mean squared error is a special case of a function called a loss function. The study of the performance, and the optimality, of estimators evaluated through loss functions is a branch of decision theory.

Definition 7.15: Action Space

After the data $\mathbf{X} = \mathbf{x}$ is observed, where $X \sim f(\mathbf{x}|\theta)$, $\theta \in \Theta$, a decision regarding θ is made. The set of allowable decisions is the action space, denoted by $\mathcal{A}$. Often in point estimation problems $\mathcal{A} = \Theta$, the parameter space, but this will change in other problems.

Definition 7.16: Loss Function

The loss function in a point estimation problem reflects the fact that if an action a is close to θ, then the decision a is reasonable and little loss is incurred. If a is far from θ, then a large loss is incurred. The loss function is a nonnegative function that generally increases as the distance between a and θ increases.

Definition 7.17: Absolute Error Loss & Squared Error Loss

Let $\theta \in \mathbb{R}$. The absolute error loss is defined as

$$L(\theta, a) = |a - \theta| \tag{7.23}$$

and the squared error loss is defined as

$$L(\theta, a) = (a - \theta)^2 \tag{7.24}$$

Definition 7.18: Risk Function

Let L denote a loss function. In decision theoretic analysis, the quality of an estimator is quantified in its risk function; that is, for an estimator $\delta(\mathbf{x})$ of θ, the risk function, a function of θ is given by

$$R(\theta, \delta) = E_\theta\big[L(\theta, \delta(\mathbf{X}))\big]. \tag{7.25}$$

For a fixed θ, the risk function is the average loss that will be incurred if the estimator $\delta(\mathbf{x})$ is used.

Definition 7.19: Bayes Risk

Let $\pi(\theta)$ denote the prior distribution of θ and $R(\theta, \delta)$ denote the risk function. We can also use a Bayesian approach to the problem of loss function optimality. In a Bayesian analysis we would use this prior distribution to compute an average risk

$$\int_\Theta R(\theta, \delta)\pi(\theta)d\theta, \tag{7.26}$$

known as Bayes Risk. The estimator that yields the smallest value of the Bayes risk is called the Bayes rule with respect to a prior π and we often denote it as δ^π.

Definition 7.20: Posterior Expected Loss

Let $\mathbf{X} \sim f(\mathbf{x}|\theta)$ and $\theta \sim \pi$. The Bayes risk of a decision rule δ can be written as

$$\int_\Theta R(\theta, \delta)\pi(\theta)d\theta = \int_\Theta \left(\int_\mathcal{X} L(\theta, \delta(\mathbf{x}))f(\mathbf{x}|\theta)d\mathbf{x} \right)\pi(\theta)d\theta = \int_\mathcal{X} \left[\int_\Theta L(\theta, \delta(\mathbf{x}))\pi(\theta|\mathbf{x})d\theta \right] m(\mathbf{x})d\mathbf{x}, \tag{7.27}$$

where the second equality came about due to $f(\mathbf{x}|\theta)\pi(\theta) = \pi(\theta|\mathbf{x})m(\mathbf{x})$ where $\pi(\theta|\mathbf{x})$ is the posterior distribution of θ and $m(\mathbf{x})$ is the marginal distribution of $\mathbf{X}$. The quantity in brackets at the right is the expected value of the loss function with respect to the posterior distribution, called the posterior expected loss. Hence, for each $\mathbf{x}$, if we choose the action $\delta(\mathbf{x})$ to minimize the posterior expected loss, we will minimize the Bayes risk.

7.3 Miscellanea

Theorem 7.9: Lehmann-Scheffe

Unbiased estimators based on complete sufficient statistics are unique.

Theorem 7.10

If the expected complete-data log likelihood $E[\log L(\theta|\mathbf{y}, \mathbf{x})|\theta', \mathbf{y}]$ is continuous in both θ and θ', then all limit points of an EM sequence $\{\hat{\theta}^{(r)}\}$ are stationary points of $L(\theta|\mathbf{y})$, and $L(\hat{\theta}^{(r)}|\mathbf{y})$ converges monotonically to $L(\hat{\theta}|\mathbf{y})$ for some stationary point $\hat{\theta}$.

Chapter 8

Hypothesis Testing

8.1 Introduction

Definition 8.1: Hypothesis Testing

A hypothesis is a statement about a population parameter.

Definition 8.2: Null & Alternative Hypothesis

The two complementary hypothesis in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis. They are denoted by H_0 and H_1, respectively.

If θ denotes a population parameter, the general format of the null and alternative hypotheses is $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$, where Θ_0 is some subset of the parameter space and Θ_0^c is its complement. For instance, if θ denotes the average change in a patient's blood pressure after taking a drug, an experimenter might be interested in testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. The null hypothesis states that, on the average, the drug has no effect on blood pressure, and the alternative hypothesis states that there is some effect. This common situation, in which H_0 states that a treatment has no effect, has lead to the term "*null*" hypothesis.

Definition 8.3: Hypothesis Testing & Rejection Regions

A hypothesis testing procedure or hypothesis test is a rule that specifies:

1. *For which sample values the decision is made to accept H_0 as true.*
2. *For which sample values H_0 is rejected and H_1 is accepted as true.*

The subset of the sample space for which H_0 will be rejected is called the rejection region or critical region. The complement of the rejection region is called the acceptance region.

8.2 Methods of Finding Tests

8.2.1 Likelihood Ratio Tests

Definition 8.4: Likelihood Ratio Test (LRT) Statistic

The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(\mathbf{x}) = \frac{\sup\limits_{\Theta_0} L(\theta|\mathbf{x})}{\sup\limits_{\Theta} L(\theta|\mathbf{x})}. \tag{8.1}$$

A likelihood ratio test (LRT) is any test that has a rejection region of the form $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$, where c is any number satisfying $0 \leq c \leq 1$.

We note that $0 \leq \lambda(\mathbf{x}) \leq 1$ as $\Theta_0 \subset \Theta$. If $\lambda(\mathbf{x}) < 1$, then it would suggest that the data is "better" explained by H_1. This forms the rationale behind the construction of the rejection region $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$ with c satisfying $0 \leq c \leq 1$. Since we want to be quite confident that H_1 is indeed true, we can set c to a relatively small value "near" zero as a value "near" 1 may have been anomalous. So long as $c \in [0, 1]$, do we consider such a test to be an LRT.

The connection to MLE's is straightforward, Suppose that $\hat{\theta} \in \Theta$ is the MLE that maximizes $L(\theta|\mathbf{x})$ over the entire parameter space and $\hat{\theta}_0 \in \Theta_0$ is the MLE that maximizes $L(\theta|\mathbf{x})$ over the restricted parameter space. Then, the LRT statistic is

$$\lambda(\mathbf{x}) = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})} \tag{8.2}$$

Theorem 8.1

If $T(\mathbf{X})$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(\mathbf{x})$ are the LRT statistics based on T and $\mathbf{X}$, respectively, then $\lambda^*(T(\mathbf{x})) = \lambda(\mathbf{x})$ for every $\mathbf{x}$ in the sample space.

8.2.2 Bayesian Tests

The Bayesian paradigm prescribes that the sample information be combined with the prior information using Bayes' Theorem to obtain the posterior distribution $\pi(\theta|\mathbf{x})$. In a hypothesis testing problem, the posterior distribution may be used to calculate the probabilities that H_0 and H_1 are true. Remember, $\pi(\theta|\mathbf{x})$ is a probability distribution for a random variable. Hence, the posterior probabilities $P(\theta \in \Theta_0|\mathbf{x}) = P(H_0 \text{ is true}|\mathbf{x})$ and $P(\theta \in \Theta_0^c|\mathbf{x}) = P(H_1 \text{ is true}|\mathbf{x})$ may be computed.

However, the probabilities $P(H_0 \text{ is true}|\mathbf{x})$ and $P(H_1 \text{ is true}|\mathbf{x})$ are not meaningful to the classical statistician. The classical statistician consider θ to be a fixed number. Consequently, a hypothesis is *either true or false*. If $\theta \in \Theta_0$, then $P(H_0 \text{ is true}|\mathbf{x}) = 1$ and $P(H_1 \text{ is true}|\mathbf{x}) = 0$ for all values of $\mathbf{x}$.

One way a Bayesian hypothesis tester may choose to use the posterior distribution is to decide to accept H_0 as true if $P(\theta \in \Theta_0|\mathbf{X}) \leq P(\theta \in \Theta_0^c|\mathbf{X})$ and to reject H_0 otherwise.

8.2.3 Union-Intersection and Intersection-Union Tests

In some situations, tests for complicated null hypotheses can be developed from tests for simpler null hypotheses.

Definition 8.5: Union-Intersection Method

The union-intersection method of test construction might be useful when the null hypothesis is conveniently expressed as an intersection, say

$$H_0 : \theta \in \cap_{\gamma\in\Gamma}\Theta_\gamma. \tag{8.3}$$

Here Γ is an arbitrary index set that may be finite or infinite, depending on the problem. Suppose tests are available for each of the problems of testing $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$. If the rejection region for the test of $H_{0\gamma}$ is

$$\{\mathbf{x} : T_\gamma(\mathbf{x}) \in R_\gamma\}. \tag{8.4}$$

Then the rejection region for the union-intersection test is

$$\cup_{\gamma\in\Gamma} \{\mathbf{x} : T_\gamma(\mathbf{x}) \in R_\gamma\}. \tag{8.5}$$

If any one of the hypotheses $H_{0\gamma}$ is rejected, then H_0, which is true only if $H_{0\gamma}$ is true for every γ, must also be rejected.

Definition 8.6: Intersection-Union Method

The intersection-union method may be useful if the null hypothesis is conveniently expressed as a union. Suppose that our null hypothesis can be expressed as

$$H_0 : \theta \in \cup_{\gamma\in\Gamma}\Theta_\gamma. \tag{8.6}$$

Suppose that for each $\gamma \in \Gamma, \{\mathbf{x} : T_\gamma(\mathbf{x}) \in R_\gamma\}$ is the rejection region for a test of $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$. Then the rejection region for the intersection-union test of H_0 versus H_1 is

$$\cap_{\gamma\in\Gamma} \{\mathbf{x} : T_\gamma(\mathbf{x}) \in R_\gamma\}. \tag{8.7}$$

From (8.6), we see that H_0 is false if and only if all of the $H_{0\gamma}$ are false, so H_0 can be rejected if and only if each of the individual hypotheses $H_{0\gamma}$ can be rejected.

8.3 Methods of Evaluating Tests

When deciding to accept or reject the null hypothesis H_0, an experimenter might be making a mistake. Usually, hypothesis tests are evaluated and compared through their probabilities of making mistakes.

8.3.1 Error Probabilities and the Power Function

Definition 8.7: Type *I* Error and Type *II* Error

Consider a hypothesis test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. If $\theta \in \Theta_0$ but the hypothesis test incorrectly decides to reject H_0, then the test has made a Type I Error. If, on the other hand, $\theta \in \Theta_0^c$ but the test decides to accept H_0, a Type II Error has been made.

Truth	*Decision*	
	Accept H_0	***Reject*** H_0
H_0	*Correct Decision*	*Type I Error*
H_1	*Type II Error*	*Correct Decision*

Suppose that R denotes the rejection region for a test. Then for $\theta \in \Theta_0$, the test will make a mistake if $\mathbf{x} \in R$, so the probability of a Type I Error is $P_\theta(\mathbf{X} \in R)$. For $\theta \in \Theta_0^c$, the probability of a Type II Error is

$P_\theta(\mathbf{X} \in R^c)$ which is related to each other by $P_\theta(\mathbf{X} \in R^c) = 1 - P_\theta(\mathbf{X} \in R)$. Hence, $P_\theta(\mathbf{X} \in R)$ contains all the information about the test with rejection region R, compactly summarized by

$$P_\theta(\mathbf{X} \in R) = \begin{cases} \text{probability of a Type I Error} & \text{if } \theta \in \Theta_0 \\ \text{one minus the probability of a Type II Error} & \text{if } \theta \in \Theta_0^c \end{cases} \tag{8.8}$$

Definition 8.8: Power Function

The power function of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(\mathbf{X} \in R)$.

The ideal power function is 0 for all $\theta \in \Theta_0$ and 1 for all $\theta \in \Theta_0^c$. Qualitatively, a good test has power function near 1 for most $\theta \in \Theta_0^c$ and near 0 for most $\theta \in \Theta_0$.

For a fixed sample size, it is usually impossible to make both types of error probabilities arbitrarily small. In searching for a good test, it is common to restrict consideration to tests that control the Type I Error probability at a specified level.

Definition 8.9: Size α Test

For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size α test if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.

Definition 8.10: Level α Test

For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a level α test if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.

Some authors do not make the distinction between the terms *size* and *level* that we have made, and sometimes these terms are used interchangeably. Experimenters commonly specify the level of the test they wish to use, with typical choices being $\alpha = 0.01, 0.05$ and 0.10.

Definition 8.11: Unbiased Test

A test with power function $\beta(\theta)$ is unbiased if $\beta(\theta') \geq \beta(\theta'')$ for every $\theta' \in \Theta_0^c$ and $\theta'' \in \Theta_0$.

8.3.2 Most Powerful Tests

Definition 8.12: Uniformly Most Powerful (UMP) Class Test

Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test in class $\mathcal{C}$, with power function $\beta(\theta)$, is a uniformly most powerful (UMP) class $\mathcal{C}$ test if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta_0^c$ and every $\beta'(\theta)$ that is a power function of a test in class $\mathcal{C}$.

In this section, the class $\mathcal{C}$ will be the class of *all level α tests* The test described in the above definition is then called a UMP level α test.

The requirements in the above definition are so strong that UMP tests do not exist in many realistic problems. The following theorem describes which tests are UMP level α tests in the situation where the null and alternative hypotheses both consist of only one probability distribution for the sample (i.e when H_0 and H_1 are simple hypotheses).

Theorem 8.2: Neyman-Pearson Lemma

Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the PDF or PMF corresponding to θ_i is $f(\mathbf{x}|\theta_i)$, $i = 0, 1$, using a test with rejection region R that satisfies

$$\begin{aligned} &\mathbf{x} \in R \text{ if } f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0) \\ &\text{and} \\ &\mathbf{x} \in R^c \text{ if } f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0), \end{aligned} \tag{8.9}$$

for some $k \geq 0$, and

$$\alpha = P_{\theta_0}(\mathbf{X} \in R) \tag{8.10}$$

Then

- **a.** *(Sufficiency) Any test that satisfies (8.12) and (8.13) is a UMP level α test.*
- **b.** *(Necessity) If there exists a test satisfying (8.12) and (8.13) with $k > 0$, then every UMP level α test is a size α test (satisfies (8.13)) and every UMP level α test satisfies (8.12) except perhaps on a set A satisfying*

$$P_{\theta_0}(\mathbf{X} \in A) = P_{\theta_1}(\mathbf{X} \in A) = 0. \tag{8.11}$$

Corollary 8.1

Consider the hypothesis problem posed in Theorem 8.2. Suppose that $T(\mathbf{X})$ is a sufficient statistic for θ and $g(t|\theta_i)$ is the PDF or PMF of T corresponding to $\theta_i, i = 0, 1$. Then any test based on T with rejection region S (a subset of the sample space of T) is a UMP level α test if it satisfies

$$\begin{aligned} &t \in S \text{ if } g(t|\theta_1) > kg(t|\theta_0) \\ &\text{and} \\ &t \in S^c \text{ if } g(t|\theta_1) < kg(t|\theta_0), \end{aligned} \tag{8.12}$$

for some $k \geq 0$, and

$$\alpha = P_{\theta_0}(T \in S) \tag{8.13}$$

Definition 8.13: Monotone Likelihood Ratio

A family of PDFs or PMFs $\{g(t|\theta) : \theta \in \Theta\}$ for a univariate random variable T with real-valued parameter θ has a monotone likelihood ratio (MLR) if, for every $\theta_2 > \theta_1$, $g(t|\theta_2)/g(t|\theta_1)$ is a monotone (nonincreasing or nondecreasing) function of t on $\{t : g(t|\theta_1) > 0 \text{ or } g(t|\theta_2) > 0\}$. Note that $c/0$ is defined as ∞ if $c > 0$.

Theorem 8.3: Karlin-Rubin

Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose that T is a sufficient statistic for θ and the family of PDFs or PMFs $\{g(t|\theta) : \theta \in \Theta\}$ of T has an MLR. Then for any t_0, the test that rejects H_0 if and only if $T > t_0$ is a UMP level α test, where $\alpha = P_{\theta_0}(T > t_0)$

8.3.3 Sizes of Union-Intersection and Intersection-Union Tests

Theorem 8.4

Consider testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$, where $\Theta_0 = \cap_{\gamma\in\Gamma}\Theta_\gamma$ and let $\lambda_\gamma(\mathbf{x})$ be the LRT statistic for testing $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$. Define $T(\mathbf{x}) = inf_{\gamma\in\Gamma}\lambda_\gamma(\mathbf{x})$, and form the UIT with rejection region

$$\{\mathbf{x} : \lambda_\gamma(\mathbf{x}) < c \text{ for some } \gamma \in \Gamma\} = \{\mathbf{x} : T(\mathbf{x}) < c\} \tag{8.14}$$

Also consider the usual LRT with rejection region $\{\mathbf{x} : \lambda(\mathbf{x}) < c\}$. Then

1. *$T(\mathbf{x}) \geq \lambda(\mathbf{x})$ for every $\mathbf{x}$;*
2. *If $\beta_T(\theta)$ and $\beta_\lambda(\theta)$ are the power functions for the tests based on T and λ, respectively, then $\beta_T(\theta) \leq \beta_\lambda(\theta)$ for every $\theta \in \Theta$;*
3. *If the LRT is a level α test, then the UIT is a level α test.*

Theorem 8.5

Let α_γ be the size of the test of $H_{0\gamma}$ with rejection region R_γ. Then the IUT with rejection region $R = \cap_{\gamma\in\Gamma}R_\gamma$ is a level $\alpha = sup_{\gamma\in\Gamma}\alpha_\gamma$ test.

Theorem 8.6

Consider testing $H_0 : \theta \in \cup_{j=1}^k\Theta_j$ where k is a finite positive integer. For each $j = 1, ..., k$, let R_j be the rejection region of a level α test of H_{0j}. Suppose that for some $i = 1, ..., k$, there exists a sequence of parameter points, $\theta_l \in \Theta_i, l = 1, 2, ...$, such that

(i) $\lim_{l\to\infty} P_{\theta_l}(\mathbf{X} \in R_i) = \alpha$,

(ii) for each $j = 1, ..., k$, $j \neq i$, $\lim_{l\to\infty} P_{\theta_l}(\mathbf{X} \in R_j) = 1$.

Then, the IUT with rejection region $R = \cap_{j=1}^k R_j$ is a size α test.

8.3.4 p-Values

After a hypothesis test is done, the conclusions must be reported in some statistically meaningful way. One method of reporting the results of a hypothesis test is to report the size, α, of the test used and the decision to reject H_0 or accept H_0. Another way of reporting the results of a hypothesis test is to report the value of a certain kind of test statistic called a *p-value*.

Definition 8.14: p-value

A p-value $p(\mathbf{X})$ is a test statistic satisfying $0 \leq p(\mathbf{x}) \leq 1$ for every sample point $\mathbf{x}$. Small values of $p(\mathbf{X})$ give evidence that H_1 is true. A p-value is valid if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$,

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha \tag{8.15}$$

An advantage to reporting a test result via a p-value is that each reader can choose the α he or she considers appropriate and then can compare the reported $p(\mathbf{x})$ to α and know whether these data lead to acceptance or rejection of H_0. A p-value report the results of a test on a more continuous scale, rather than just the dichotomous decision "*Accept H_0*" or "*Reject H_0*".

Theorem 8.7

Let $W(\mathbf{X})$ be a test statistic such that large values of W give evidence that H_1 is true[a]*. For each sample point $\mathbf{x}$, define*

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta\big(W(\mathbf{X}) \geq W(\mathbf{x})\big). \tag{8.16}$$

Then, $p(\mathbf{X})$ is a valid p-value.

[a] I take this to mean that there exists an $n \in \mathbb{R}$ such that if $W(\mathbf{X}) \geq n$, we conclude that H_1 is true.

8.3.5 Loss Function Optimality

In a hypothesis testing problem, only two actions are allowable, "*accept* H_0" or "*reject* H_0". These two actions could be denoted by a_0 and a_1 respectively. The action space in hypothesis testing is the two-point set $\mathcal{A} = \{a_0, a_1\}$. A decision rule $\delta(\mathbf{x})$ (a hypothesis test) is a function on $\mathcal{X}$ that takes on only two values, a_0 and a_1. The set $\{\mathbf{x} : \delta(\mathbf{x}) = a_0\}$ is the acceptance region for the test and the set $\{\mathbf{x} : \delta(\mathbf{x}) = a_1\}$ is the rejection region.

Definition 8.15: 0-1 Loss

The simplest kind of loss in a testing problem is called $0-1$ loss and is defined by

$$L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ 1 & \theta \in \Theta_0^c \end{cases} \qquad \text{and} \qquad L(\theta, a_1) = \begin{cases} 1 & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c \end{cases} \tag{8.17}$$

Definition 8.16: Generalized 0-1 Loss

A slightly more realistic loss that gives different costs to the two types of error is called the generalized $0-1$ loss:

$$L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ c_{II} & \theta \in \Theta_0^c \end{cases} \qquad \text{and} \qquad L(\theta, a_1) = \begin{cases} c_I & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c \end{cases}. \tag{8.18}$$

In this loss, c_I is the cost of a Type I error, the error of falsely rejecting H_0, and c_{II} is the cost of a Type II Error, the error of falsely accepting H_0.

The *0-1 loss* only judges whether a decision is right or wrong. It may be the case that some wrong decisions are more serious than others and the loss function should reflect this.

Alphabetical Index

A

Ancillary Statistic 55

B

Beta Distribution 20
Bias 64
Bias-Variance Decomposition 64
Binomial Distribution 16

C

Cauchy Distribution 20
Central Limit Theorem 47
Central Moment 11
Chi-Squared Distribution 19
Complete Statistic 55
Conditionality Principle 56
Converges Almost Surely 47
Converges in Distribution 47
Converges in Probability 46
Correlation 33
Covariance 32
Cumulative Distribution Function 6

D

Double Exponential Distribution 21

E

Evidence Function 56
Expected Value 10
Experiment 56
Exponential Distribution 19

F

Formal Likelihood Principle 57
Formal Sufficiency Principle 56

G

Gamma Distribution 18
Geometric Distribution 17

H

Hypergeometric Distribution 15

I

Independent Random Variables 30
Induced Probability Function 6

L

Likelihood Function 56
Likelihood Principle 56
Likelihood Ratio Test 70
Lognormal Distribution 21

M

Mean Squared Error 63
Method of Moments 61
Minimal Sufficient Statistic 54
Moment 11
Moment Generating Function 11

N

Negative Binomial Distribution 17
Normal Distribution 20

O

Order Statistics 44

P

Poisson Distribution 16
Probability Mass Function 7

R

Random Variable 6

S

Sample Mean 40
Sample Standard Deviation 40
Sample Variance 40
SLLN 47
Standard Deviation 11
Sufficient Statistic 54

U

Uniform Distribution
- Continuous 18
- Discrete 15

V

Variance 11

W

Weibull Distribution 19
WLLN 46