MATH 541A, GRADUATE STATISTICS, SPRING 2023

STEVEN HEILMAN

Abstract. These lecture notes are based upon the textbooks Cassella and Berger, Statistical Inference and Keener, Theoretical Statistics.

Contents

1. Review of Probability Theory 2
1.1. Random Variables, Expectation 2
1.2. Examples of Random Variables 5
1.3. Expected Value 10
1.4. Joint PDFs 12
1.5. Conditional Probability and Conditional Expectation 16
1.6. Functions of Random Variables 20
1.7. Inequalities 22
1.8. Independent Sums and Convolution 24
1.9. Additional Comments 26
2. Limit Theorems 27
2.1. Modes of Convergence 27
2.2. Limit Theorems 28
2.3. Additional Comments 33
3. Exponential Families 36
3.1. Differential Identities 38
3.2. Additional Comments 42
4. Random Samples 42
4.1. Sampling from the Normal 43
4.2. The Delta Method 47
4.3. Simulation of Random Variables 49
4.4. Additional Comments 52
5. Data Reduction 52
5.1. Sufficient Statistics 53
5.2. Ancillary Statistics 58
5.3. Complete Statistics 59
5.4. Additional Comments 63
6. Point Estimation 64
6.1. Heuristic Principles for Finding Good Estimators 64
6.2. Evaluating Estimators 65
6.3. Efficiency of an Estimator 70

Date: October 25, 2023

6.4. Bayes Estimation 73
6.5. Method of Moments 74
6.6. Maximum Likelihood Estimator 76
6.7. EM Algorithm 83
6.8. Additional Comments 84
7. Resampling and Bias Reduction 85
7.1. Jackknife Resampling 86
7.2. Bootstrapping 87
8. Some Concentration of Measure 89
8.1. Concentration for Independent Sums 89
8.2. Concentration for Lipschitz Functions 91
8.3. Additional Comments 94
9. Appendix: Results from Analysis 94
10. Appendix: Convergence in Distribution, Characteristic Functions 99
11. Appendix: Moment Generating Functions 102
12. Appendix: Notation 104

1. Review of Probability Theory

1.1. Random Variables, Expectation.

Definition 1.1 (**Universal Set**). In a specific problem, we assume the existence of a sample space, or **universal set** Ω which contains all other sets. The universal set represents all possible outcomes of some random process. We sometimes call the universal set the **universe**. The universe is always assumed to be nonempty. Subsets of the sample space are sometimes called **events**.

Definition 1.2 (**Countable Set Operations**). Let $A_1, A_2, \ldots \subseteq \Omega$. We define

$$\bigcup_{i=1}^{\infty} A_i = \{x \in \Omega \colon \exists \text{ a positive integer } j \text{ such that } x \in A_j\}.$$

$$\bigcap_{i=1}^{\infty} A_i = \{x \in \Omega \colon x \in A_j, \ \forall \text{ positive integers } j\}.$$

Exercise 1.3. Prove that the set of real numbers $\mathbb{R}$ can be written as the countable union

$$\mathbb{R} = \bigcup_{j=1}^{\infty} [-j, j].$$

(Hint: you should show that the left side contains the right side, and also show that the right side contains the left side.)

Prove that the singleton set $\{0\}$ can be written as

$$\{0\} = \bigcap_{j=1}^{\infty} [-1/j, 1/j].$$

Definition 1.4 (**Disjointness**). Let A, B be sets in some universe Ω. We say that A and B are **disjoint** if $A \cap B = \emptyset$. A collection of sets $A_1, A_2, \ldots$ in Ω is said to be a **partition** of Ω if $\cup_{i=1}^{\infty} A_i = \Omega$, and if, for all $i, j \geq 1$ with $i \neq j$, we have $A_i \cap A_j = \emptyset$.

Remark 1.5. Two or three sets can be visualized with a Venn diagram, though the Venn diagram is no longer very helpful when considering more than three sets.

The following properties follow from the above definitions.

Proposition 1.6. *Let A, B, C be sets in a universe Ω.*

(i) $A \cup B = B \cup A$.
(ii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
(iii) $(A^c)^c = A$.
(iv) $A \cup \Omega = \Omega$.
(v) $A \cup (B \cup C) = (A \cup B) \cup C$.
(vi) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
(vii) $A \cap A^c = \emptyset$.
(viii) $A \cap \Omega = A$.

Exercise 1.7. Using the definitions of intersection, union and complement, prove properties (ii) and (iii). (Hint: to prove property (ii), it may be helpful to first draw a Venn diagram of A, B, C. Now, let $x \in \Omega$. Consider where x could possibly be with respect to A, B, C. For example, we could have $x \in A$, $x \notin B$, $x \in C$. We could also have $x \in A$, $x \in B$, $x \notin C$. And so on. In total, there should be $2^3 = 8$ possibilities for the location of x, with respect to A, B, C. Construct a truth table which considers all eight such possibilities for each side of the purported equality $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.)

Exercise 1.8 (**De Morgan's Laws**). Let $A_1, A_2, \ldots$ be sets in some universe Ω. Then

$$\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c, \qquad \left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c.$$

Exercise 1.9. Let $A_1, A_2, \ldots$ be sets in some universe Ω. Let $B \subseteq \Omega$. Show the following generalization of Proposition 1.6(ii).

$$B \cap \left(\bigcup_{k=1}^{\infty} A_k\right) = \bigcup_{k=1}^{\infty} (A_k \cap B).$$

Exercise 1.10. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function. Show that

$$\cup_{y \in \mathbb{R}} \{x \in \mathbb{R}\colon f(x) = y\} = \mathbb{R}.$$

Also, show that the union on the left is disjoint. That is, if $y_1 \neq y_2$ and $y_1, y_2 \in \mathbb{R}$, then $\{x \in \mathbb{R}\colon f(x) = y_1\} \cap \{x \in \mathbb{R}\colon f(x) = y_2\} = \emptyset$.

Definition 1.11. A **Probability Law** (or **probability distribution**) $\mathbf{P}$ on a sample space Ω is a function whose domain is the set of all subsets of Ω, and whose range is contained in $[0, 1]$, such that

(i) For any $A \subseteq \Omega$, we have $\mathbf{P}(A) \geq 0$. (**Nonnegativity**)

(ii) For any $A, B \subseteq \Omega$ such that $A \cap B = \emptyset$, we have

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

If $A_1, A_2, \ldots \subseteq \Omega$ and $A_i \cap A_j = \emptyset$ whenever i, j are positive integers with $i \neq j$, then

$$\mathbf{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbf{P}(A_k). \qquad \textbf{(Additivity)}$$

(iii) We have $\mathbf{P}(\Omega) = 1$. (**Normalization**)

More generally, a **measure** μ satisfies properties (i) and (ii) and has a range in $[0, \infty]$.

Remark 1.12. For technical reasons, it is sometimes not possible to define a probability law on an arbitrary uncountable sample space. However, in practice, many sample spaces will be finite or countable, so this issue will not arise in many applications of statistics. Nevertheless, this is an important foundational issue in probability theory; for more on the subject, take a class on measure theory, or consult my graduate probability notes here.

Proposition 1.13 (**Properties of Probability Laws**). *Let Ω be a sample space and let $\mathbf{P}$ be a probability law on Ω. Let $A, B, C \subseteq \Omega$.*

- *If $A \subseteq B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.*
- $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.
- $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.
- $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A^c \cap B^c \cap C)$.

Let n be a positive integer. Let $A_1, \ldots, A_n \subseteq \Omega$. Then

$$\mathbf{P}\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} \mathbf{P}(A_k).$$

Proof. Let $A \subseteq B$. Then $B = (B \cap A) \cup (B \cap A^c)$, and $(B \cap A) \cap (B \cap A^c) = \emptyset$. So, using Axiom (ii) for probability laws, $B \cap A = A$, and using Axiom (i) for probability laws,

$$\mathbf{P}(B) = \mathbf{P}(B \cap A) + \mathbf{P}(B \cap A^c) = \mathbf{P}(A) + \mathbf{P}(B \cap A^c) \geq \mathbf{P}(A).$$

So, the first item is proven. We now prove the second item. Write $A = (A \setminus B) \cup (A \cap B)$ and note that $A \setminus B$ and $A \cap B$ are disjoint. Similarly, write $B = (B \setminus A) \cup (B \cap A)$ and note that $(B \setminus A)$ and $(B \cap A)$ are disjoint. Finally, we can write $A \cup B$ as the union of three disjoint sets: $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$.

So, using Axiom (ii) for probability laws twice,

$$\mathbf{P}(A) + \mathbf{P}(B) = \mathbf{P}(A \setminus B) + \mathbf{P}(A \cap B) + \mathbf{P}(B \setminus A) + \mathbf{P}(A \cap B) = \mathbf{P}(A \cup B) + \mathbf{P}(A \cap B).$$

So, the second item is proven. The third and fourth items are left to the exercises. The final inequality follows from the third item and induction on n. □

Definition 1.14 (**Random Variable**). Let Ω be a sample space. Let $\mathbf{P}$ be a probability law on Ω. A **random variable** X is a function $X \colon \Omega \to \mathbb{R}$. (Sometimes we might also consider a random variable to be a function from Ω to another set.) Let n be a positive integer. A **random vector** X is a function $X \colon \Omega \to \mathbb{R}^n$. A **discrete random variable** is a random variable whose range is either finite or countably infinite. A **probability density function** (PDF) is a function $f \colon \mathbb{R} \to [0, \infty)$ such that $\int_{-\infty}^{\infty} f(x)dx = 1$, and such that, for any

$-\infty \le a \le b \le \infty$, the integral $\int_a^b f(x)dx$ exists. A random variable X is called **continuous** if there exists a probability density function f such that, for any $-\infty \le a \le b \le \infty$, we have

$$\mathbf{P}(a \le X \le b) = \int_a^b f(x)dx.$$

When this equality holds, we call f the **probability density function of** X.

Let X be any random variable. We then define the **cumulative distribution function** (CDF) $F\colon \mathbb{R} \to [0,1]$ of X by

$$F(x) := \mathbf{P}(X \le x), \qquad \forall\, x \in \mathbb{R}.$$

We say two random variables X, Y are **identically distributed** if they have the same CDF.

Remark 1.15. There is another foundational issue here for uncountable sample spaces which we will not discuss further. It suffices to say that the definition of a random variable should have an extra condition, which is not needed for finite or countable sample spaces; for more on the subject, take a class on measure theory, or consult my graduate probability notes here.

Definition 1.16 (**Probability Mass Function**)**.** Let X be a discrete random variable on a sample space Ω, so that $X\colon \Omega \to \mathbb{R}$. The **probability mass function** (or PMF) of X, denote $f_X\colon \mathbb{R} \to [0,1]$ is defined by

$$f_X(x) = \mathbf{P}(X = x) = \mathbf{P}(\{X = x\}) = \mathbf{P}(\{\omega \in \Omega\colon X(\omega) = x\}), \qquad x \in \mathbb{R}.$$

Definition 1.17 (**Independence**)**.** Let $A_1, A_2, \ldots$ be subsets of a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. We say that $A_1, A_2, \ldots$ are **independent** if, for any finite subset S of $\{1, 2, \ldots\}$, we have

$$\mathbf{P}\left(\cap_{i \in S} A_i\right) = \prod_{i \in S} \mathbf{P}(A_i).$$

Let $X_1\colon \Omega \to \mathbb{R}^n, X_2\colon \Omega \to \mathbb{R}^n, \ldots$ be random variables. We say that $X_1, X_2, \ldots$ are **independent** if, for any integer $m \ge 1$ and for any $B_1, B_2, \ldots, \subseteq \mathbb{R}^n$,

$$\mathbf{P}\left(\cap_{i=1}^m \{X_i \in B_i\}\right) = \prod_{i=1}^m \mathbf{P}(X_i \in B_i).$$

Here we denoted $\{X \in B\} := \{\omega \in \Omega\colon X(\omega) \in B\}$ where $X\colon \Omega \to \mathbb{R}^n$ and $B \subseteq \mathbb{R}^n$.

1.2. **Examples of Random Variables.** We now give descriptions of some commonly encountered random variables.

Definition 1.18 (**Bernoulli Random Variable**)**.** Let $0 < p < 1$. A random variable X is called a **Bernoulli random variable with parameter** p if X has the following PMF:

$$\mathbf{P}(X = k) = \begin{cases} p & \text{, if } k = 1 \\ 1 - p & \text{, if } k = 0 \\ 0 & \text{, otherwise.} \end{cases}$$

Definition 1.19 (**Binomial Random Variable**)**.** Let $0 < p < 1$ and let n be a positive integer. A random variable X is called a **binomial random variable with parameters** n **and** p if X has the following PMF. If k is an integer with $0 \leq k \leq n$, then

$$\mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

For any other k, we have $\mathbf{P}(X = k) = 0$.

Recall that a sum of n independent Bernoulli random variables with parameter $0 < p < 1$ is a binomial random variable with parameters n and p.

Definition 1.20 (**Geometric Random Variable**)**.** Let $0 < p < 1$. A random variable X is called a **geometric random variable with parameter** p if X has the following PMF. If k is a positive integer, then

$$\mathbf{P}(X = k) = (1-p)^{k-1} p.$$

For any other k, we have $\mathbf{P}(X = k) = 0$. Note that X is the number of times that are needed to flip a biased coin in order to get a heads (if the coin has probability p of landing heads).

Definition 1.21 (**Negative Binomial Random Variable**)**.** Let $0 < p < 1$ and let n be a positive integer. A random variable X is called a **negative binomial random variable with parameters** n **and** p if X has the following PMF. If k is an integer with $n \leq k$, then

$$\mathbf{P}(X = k) = \binom{k-1}{n-1} (1-p)^{k-n} p^n.$$

For any other k, we have $\mathbf{P}(X = k) = 0$. Note that X is the number of times that are needed to flip a biased coin in order to get n heads (if the coin has probability p of landing heads). The case $n = 1$ recovers the geometric random variable.

The negative binomial is equivalently defined as $Y = X - n$, i.e. the number of tails that occur before the n^{th} heads occurs, so that for any $k \geq 0$,

$$\mathbf{P}(Y = k) = \mathbf{P}(X = k + n) = \binom{k+n-1}{n-1} (1-p)^k p^n = \binom{k+n-1}{k} (1-p)^k p^n.$$

Definition 1.22 (**Hypergeometric Random Variable**)**.** Let m, n, p be positive integers such that $m \leq p$. A random variable X is called a **hypergeometric random variable with parameters** m, n, p if X has the following PMF. If k is a positive integer with $\max(0, p + m - n) \leq k \leq \min(m, p)$, then

$$\mathbf{P}(X = k) = \frac{\binom{m}{k}\binom{n-m}{p-k}}{\binom{n}{p}}$$

For any other k, we have $\mathbf{P}(X = k) = 0$.

Suppose we have an urn containing n cubes, where m cubes are red and the remaining $n - m$ cubes are blue. We then randomly select p cubes from the urn, without replacement. Let $0 \leq k \leq m$ be an integer. Then the probability that exactly k of the selected cubes are red is given by the above distribution, since $\binom{m}{k}$ is the number of ways to select k of the (labelled) red cubes, $\binom{n-m}{p-k}$ is the number of ways to select $p-k$ of the (labelled) blue cubes, and we then divide by the total number of ways to select p cubes from all n of them.

Definition 1.23 (**Poisson Random Variable**)**.** Let $\lambda > 0$. A random variable X is called a **Poisson random variable with parameter** λ if X has the following PMF. If k is a nonnegative integer, then

$$\mathbf{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

For any other x, we have $p_X(x) = 0$.

Example 1.24. We say that a random variable X is **uniformly distributed in** $[c, d]$ when X has the following density function: $f(x) = \frac{1}{d-c}$ when $x \in [c, d]$, and $f(x) = 0$ otherwise.

Example 1.25. Let $\lambda > 0$. A random variable X is called an **exponential random variable with parameter** λ if X has the following density function: $f(x) = \lambda e^{-\lambda x}$ when $x \geq 0$, and $f(x) = 0$ otherwise.

Definition 1.26 (**Normal Random Variable**)**.** Let $\mu \in \mathbb{R}$, $\sigma > 0$. A continuous random variable X is said to be **normal** or **Gaussian** with mean μ and variance σ^2 if X has the following density function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \forall\, x \in \mathbb{R}.$$

In particular, a **standard normal** or **standard Gaussian** random variable is defined to be a normal with $\mu = 0$ and $\sigma = 1$.

Proposition 1.27 (**Poisson Approximation to the Binomial**)**.** *Let $\lambda > 0$. For each positive integer n, let $0 < p_n < 1$, and let X_n be a binomial distributed random variable with parameters n and p_n. Assume that $\lim_{n\to\infty} p_n = 0$ and $\lim_{n\to\infty} np_n = \lambda$. Then, for any nonnegative integer k, we have*

$$\lim_{n\to\infty} \mathbf{P}(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Lemma 1.28. *Let $\lambda > 0$. For each positive integer n, let $\lambda_n > 0$. Assume that $\lim_{n\to\infty} \lambda_n = \lambda$. Then*

$$\lim_{n\to\infty} \left(1 - \frac{\lambda_n}{n}\right)^n = e^{-\lambda}$$

Proof. Let log denote the natural logarithm. For any $x < 1$, define $f(x) = \log(1 - x)$. From L'Hôpital's Rule,

$$\lim_{x\to 0} \frac{f(x)}{x} = \lim_{x\to 0} f'(x) = \lim_{x\to 0} \frac{-1}{1-x} = -1. \qquad (*)$$

So, using $\lim_{n\to\infty} \lambda_n/n = 0$ we can apply $(*)$ and then $\lim_{n\to\infty} \lambda_n = \lambda$, so

$$\begin{aligned}\lim_{n\to\infty} \left(1 - \frac{\lambda_n}{n}\right)^n &= \lim_{n\to\infty} \exp\left(\log\left(1 - \frac{\lambda_n}{n}\right)^n\right) \\ &= \exp\left(\lim_{n\to\infty} \frac{\log\left(1 - \frac{\lambda_n}{n}\right)}{\lambda_n/n}\lambda_n\right) = \exp((-1)(\lambda)) = e^{-\lambda}.\end{aligned}$$

□

Proof of Proposition 1.27. For any positive integer n, let $\lambda_n = np_n$. Then $\lim_{n\to\infty} \lambda_n = \lambda$ and $\lim_{n\to\infty} \lambda_n/n = 0$. And if k is a nonnegative integer,

$$\begin{aligned}\mathbf{P}(X_n = k) &= \binom{n}{k}\left(\frac{\lambda_n}{n}\right)^k\left(1-\frac{\lambda_n}{n}\right)^{n-k} \\ &= \left(\prod_{i=1}^{k}\frac{n-i+1}{n}\right)\frac{\lambda_n^k}{k!}\left(1-\frac{\lambda_n}{n}\right)^n\left(1-\frac{\lambda_n}{n}\right)^{-k}\end{aligned}$$

So, using Lemma 1.28, $\lim_{n\to\infty} \mathbf{P}(X_n = k) = 1 \cdot \frac{\lambda^k}{k!}e^{-\lambda} \cdot 1$. □

Remark 1.29. A Poisson random variable is often used as an approximation for counting the number of some random occurrences. For example, the Poisson distribution can model the number of typos per page in a book, the number of magnetic defects in a hard drive, the number of traffic accidents in a day, etc.

Exercise 1.30. The Wheel of Fortune involves the repeated spinning of a wheel with 72 possible stopping points. We assume that each time the wheel is spun, any stopping point is equally likely. Exactly one stopping point on the wheel rewards a contestant with $\$1,000,000$. Suppose the wheel is spun 24 times. Let X be the number of times that someone wins $\$1,000,000$. Using the Poisson Approximation the Binomial, estimate the following probabilities: $\mathbf{P}(X = 0)$, $\mathbf{P}(X = 1)$, $\mathbf{P}(X = 2)$. (Hint: consider the binomial distribution with $p = 1/72$.)

Remark 1.31. The Bernoulli, binomial, geometric and Poisson random variables are all examples of the following general construction of a random variable. Let $a_0, a_1, a_2, \ldots \geq 0$ such that $\sum_{i=0}^{\infty} a_i = 1$. Then define a random variable X such that $\mathbf{P}(X = i) = a_i$ for all nonnegative integers i.

There are many other random variables we will encounter in this class as well, but these will be enough for now.

Exercise 1.32. For any $\alpha > 0$ define the **Gamma function** $\Gamma(\alpha)$ by the formula

$$\Gamma(\alpha) := \int_0^\infty x^{\alpha-1}e^{-x}dx.$$

Since $\alpha > 0$, it follows that $0 \leq \int_0^\infty x^{\alpha-1}e^{-x}dx < \infty$, so this quantity is well-defined.

Using integration by parts, show that for any $\alpha > 0$, we have the recursion

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha).$$

Since $\Gamma(1) = 1$, conclude by an inductive argument that, for any positive integer n,

$$\Gamma(n+1) = n!.$$

In this way, the Gamma function extends the definition of the factorial to any positive real number.

Definition 1.33 (Gamma Distribution). Let $\alpha, \beta > 0$. Define the **gamma distribution with parameters** (α, β) to be the random variable with the probability density function

$$f(x) := \begin{cases} \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0. \end{cases}$$

By changing variables, note that

$$P(X/\beta < t) = \mathbf{P}(X < t\beta) = \int_0^{t\beta} \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} dx = \int_0^t \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)} dy.$$

That is, X/β has the gamma distribution with parameters $(\alpha, 1)$. Also, choosing $t = \infty$ shows that the integral of the density function is one on $(-\infty, \infty)$.

For example, if $\alpha = p/2$ where p is a positive integer and $\beta = 2$, we get the **chi squared distribution** with p degrees of freedom:

$$f(x) := \begin{cases} \frac{x^{p/2-1}e^{-x/2}}{2^{p/2}\Gamma(p/2)}, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0. \end{cases}$$

This distribution can be defined as the distribution of sum of the squares of p independent standard Gaussian random variables. See Example 1.109 below for a derivation of this fact when $p = 1$ or $p = 2$.

Definition 1.34 (**Beta Distribution**)**.** Let $\alpha, \beta > 0$. Define the **beta distribution with parameters** (α, β) to be the random variable with the probability density function

$$f(x) := \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & \text{if } 0 < x < 1 \\ 0, & \text{if } x \notin [0,1]. \end{cases}$$

Here $B(\alpha, \beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}$.

It can be shown that $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. The quickest proof first switches to (squared) polar coordinates so that $x = r\cos^2\theta$, $y = r\sin^2\theta$. Then the Jacobian determinant is

$$\det \begin{pmatrix} \cos^2\theta & -2r\cos\theta\sin\theta \\ \sin^2\theta & 2r\sin\theta\cos\theta \end{pmatrix} = 2r\sin\theta\cos\theta.$$

Using this change of variables, we get

$$\begin{aligned} \Gamma(\alpha)\Gamma(\beta) &= \int_0^\infty \int_0^\infty x^{\alpha-1}e^{-x}y^{\beta-1}e^{-y}dxdy \\ &= \int_0^\infty \int_0^{\pi/2} 2r^{\alpha+\beta-1}e^{-r(\cos^2\theta+\sin^2\theta)}\cos^{2\alpha-1}\theta\sin^{2\beta-1}\theta d\theta dr \\ &= 2\int_0^\infty r^{\alpha+\beta-1}e^{-r}dr \int_0^{\pi/2} \cos^{2\alpha-1}\theta\sin^{2\beta-1}\theta d\theta \\ &= \Gamma(\alpha+\beta)\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \Gamma(\alpha+\beta)B(\alpha,\beta). \end{aligned}$$

In the last line, we changed variables by $t = \cos^2\theta$, so that $dt = -2\cos\theta\sin\theta d\theta$.

Definition 1.35 (**Cauchy Distribution**)**.** Define the (centered) **Cauchy distribution** to be the random variable with the probability density function

$$f(x) := \frac{1}{\pi}\frac{1}{1+x^2}, \qquad \forall\, x \in \mathbb{R}.$$

Note that $\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1}{1+x^2}dx = \frac{1}{\pi}\tan^{-1}(x)|_{x=-\infty}^{x=\infty} = 1$. Also, from Remark 1.40, note that $\mathbf{E}\,|X| = 2\int_0^{\infty}\frac{x}{\pi(x^2+1)}dx = \infty$, so $\mathbf{E}X$ does not exist when X is a Cauchy distributed random variable.

1.3. **Expected Value.**

Definition 1.36 (**Indicator Function**)**.** Let $A \subseteq \Omega$ be a set. We define the **indicator function of** A, denoted $1_A\colon \Omega \to \mathbb{R}$ so that $1_A(\omega) = 0$ if $\omega \notin A$, and $1_A(\omega) = 1$ if $\omega \in A$.

Definition 1.37 (**Expected Value**)**.** Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a random variable on Ω. Assume that $X\colon \Omega \to [0,\infty)$. We define the **expected value** of X, denoted $\mathbf{E}(X)$, by

$$\mathbf{E}(X) = \int_0^{\infty}\mathbf{P}(X > t)dt.$$

In analytic notation, $\mathbf{E}X = \int_{\Omega}X(\omega)d\mathbf{P}(\omega)$. More generally, if $g\colon [0,\infty) \to [0,\infty)$ is a differentiable function such that g' is continuous and $g(0) = 0$, we define

$$\mathbf{E}g(X) = \int_0^{\infty}g'(t)\mathbf{P}(X > t)dt.$$

In particular, taking $g(t) = t^n$ for any positive integer n, for any $t \geq 0$, we have

$$\mathbf{E}X^n = \int_0^{\infty}nt^{n-1}\mathbf{P}(X > t)dt.$$

For a general random variable X, if $\mathbf{E}\max(X,0) < \infty$ and if $\mathbf{E}\max(-X,0) < \infty$, we then define $\mathbf{E}(X) = \mathbf{E}\max(X,0) - \mathbf{E}\max(-X,0)$. Otherwise, we say that $\mathbf{E}(X)$ is undefined.

Remark 1.38. If we assume that the expected value and the integral on $\mathbb{R}$ can be commuted, then the following derivation of the formula for $\mathbf{E}g(X)$ can be given. From the Fundamental Theorem of Calculus, we have

$$g(X) = \int_0^X g'(t)dt = \int_0^{\infty}g'(t)1_{\{X>t\}}dt.$$

Therefore, $\mathbf{E}g(X) = \mathbf{E}\int_0^{\infty}g'(t)1_{\{X>t\}}dt = \int_0^{\infty}g'(t)\mathbf{E}1_{\{X>t\}}dt = \int_0^{\infty}g'(t)\mathbf{P}(X > t)dt$.

Remark 1.39. If X only takes positive integer values, then for any $t > 0$, if k is an integer such that $k - 1 < t \leq k$, then $\mathbf{P}(X > t) = \mathbf{P}(X \geq k)$, so

$$\mathbf{E}(X) = \int_0^{\infty}\mathbf{P}(X > t)dt = \sum_{k=1}^{\infty}\int_{k-1}^{k}\mathbf{P}(X > t)dt = \sum_{k=1}^{\infty}\mathbf{P}(X \geq k) = \sum_{k=0}^{\infty}\mathbf{P}(X > k).$$

Also, using Fubini's Theorem 1.80 to rearrange the sum, we can arrive at

$$\begin{aligned}\mathbf{E}(X) &= \sum_{k=0}^{\infty}\mathbf{P}(X > k) = \sum_{k=0}^{\infty}\sum_{j=k+1}^{\infty}\mathbf{P}(X = j) = \sum_{0\leq k<j\leq\infty}\mathbf{P}(X = j)\\ &= \sum_{j=1}^{\infty}\sum_{k=0}^{j-1}\mathbf{P}(X = j) = \sum_{j=1}^{\infty}j\mathbf{P}(X = j).\end{aligned}$$

Remark 1.40. If X is positive with density function f that is continuous, then recall that $(d/dt)\mathbf{P}(X \leq t) = f(t)$ for all $t \in \mathbb{R}$. Since $\mathbf{P}(X > t) = 1 - \mathbf{P}(X \leq t)$, we then have $(d/dt)\mathbf{P}(X > t) = -f(t)$. So, we can recover the usual formula for expected value by integrating by parts (assuming $g(0) = 0$ and $|g(t)| \leq 1$ for all $t \geq 0$):

$$\mathbf{E}g(X) = \int_0^\infty g'(t)\mathbf{P}(X > t)dt = -\int_0^\infty g(t)\frac{d}{dt}\mathbf{P}(X > t)dt = \int_0^\infty g(t)f(t)dt.$$

Exercise 1.41 (**Stein Identity**)**.** Let X be a standard Gaussian random variable, so that X has density $x \mapsto e^{-x^2/2}/\sqrt{2\pi}$, $\forall\ x \in \mathbb{R}$. Let $g\colon \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function such that g and g' have polynomial volume growth. That is, $\exists\ a, b > 0$ such that $|g(x)|, |g'(x)| \leq a(1 + |x|)^b$, $\forall\ x \in \mathbb{R}$. Prove the **Stein identity**

$$\mathbf{E}Xg(X) = \mathbf{E}g'(X).$$

Using this identity, recursively compute $\mathbf{E}X^k$ for any positive integer k.

Alternatively, for any $t > 0$, show that $\mathbf{E}e^{tX} = e^{t^2/2}$, i.e. compute the **moment generating function** of X. Then, using $\frac{d^k}{dt^k}|_{t=0}\mathbf{E}e^{tX} = \mathbf{E}X^k$ and using the power series expansion of the exponential, compute $\mathbf{E}X^k$ directly from the identity $\mathbf{E}e^{tX} = e^{t^2/2}$.

Theorem 1.42 (**Fundamental Theorem of Calculus**)**.** *Let f be a probability density function. Then the function $g(t) = \int_{-\infty}^t f(x)dx$ is continuous at any $t \in \mathbb{R}$. Also, if f is continuous at a point x, then g is differentiable at $t = x$, and $g'(x) = f(x)$.*

Proposition 1.43. *Let $X_1, \ldots, X_n$ be random variables. Then*

$$\mathbf{E}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbf{E}(X_i).$$

Unfortunately the above property is not obvious from our definition of expected value.

Definition 1.44 (**Variance**)**.** Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a random variable on Ω. We define the **variance** of X, denoted $\mathrm{var}(X)$, by

$$\mathrm{var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \mathbf{E}X^2 - (\mathbf{E}X)^2.$$

We define the **standard deviation** of X, denoted σ_X, by

$$\sigma_X = \sqrt{\mathrm{var}(X)}.$$

Proposition 1.45. *Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a random variable on Ω. Let a, b be constants. Then*

$$\mathrm{var}(aX + b) = a^2\mathrm{var}(X).$$

We will review conditional expectation later on in the notes.

Exercise 1.46 (**Inclusion-Exclusion Formula**)**.** Let $A_1, \ldots, A_n \subseteq \Omega$ be events. Then:

$$\mathbf{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbf{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathbf{P}(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} \mathbf{P}(A_i \cap A_j \cap A_k)$$
$$\cdots + (-1)^{n+1}\mathbf{P}(A_1 \cap \cdots \cap A_n).$$

To prove this formula, show that $1_{\cup_{i=1}^n A_i} = 1 - \prod_{i=1}^n (1 - 1_{A_i})$ and then take expected values of both sides.

1.4. **Joint PDFs.**

Definition 1.47 (**Joint Probability Density Function, Two Variables**). A **joint probability density function (PDF)** for two random variables is a function $f\colon \mathbb{R}^2 \to [0,\infty)$ such that $\iint_{\mathbb{R}^2} f(x,y)dxdy = 1$, and such that, for any $-\infty \le a < b \le \infty$ and $-\infty \le c < d \le \infty$, the integral $\int_{y=c}^{y=d}\int_{x=a}^{x=b} f_{X,Y}(x,y)dxdy$ exists.

Definition 1.48. Let X, Y be two continuous random variables on a sample space Ω. We say that X and Y are **jointly continuous** with **joint PDF** $f_{X,Y}\colon \mathbb{R}^2 \to [0,\infty)$ if, for any subset $A \subseteq \mathbb{R}^2$, we have

$$\mathbf{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)dxdy.$$

In particular, choosing $A = [a,b] \times [c,d]$ with $-\infty \le a < b \le \infty$ and $-\infty \le c < d \le \infty$, we have

$$\mathbf{P}(a \le X \le b, c \le Y \le d) = \int_{y=c}^{y=d}\int_{x=a}^{x=b} f_{X,Y}(x,y)dxdy.$$

We define the **marginal PDF** f_X of X by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \qquad \forall\, x \in \mathbb{R}.$$

We define the **marginal PDF** f_Y of Y by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx, \qquad \forall\, y \in \mathbb{R}.$$

Note that

$$\mathbf{P}(c \le Y \le d) = \mathbf{P}(-\infty \le X \le \infty, c \le Y \le d) = \int_{y=c}^{y=d}\int_{x=-\infty}^{x=\infty} f_{X,Y}(x,y)dxdy.$$

Comparing this formula with Definition 1.14, we see that the marginal PDF of Y is exactly the PDF of Y. Similarly, the marginal PDF of X is the PDF of X.

Example 1.49. Suppose X and Y have a joint PDF so that

$$\mathbf{P}((X,Y) \in A) = \frac{1}{2\pi}\iint_A e^{-(x^2+y^2)/2}dxdy.$$

That is, we can think of X as the x-coordinate of a randomly thrown dart, and we can think of Y as the y-coordinate of a randomly thrown dart on the infinite dartboard $\mathbb{R}^2$.

In this case, the marginals are both standard Gaussians:

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\int_{-\infty}^{\infty} e^{-y^2/2}\frac{dy}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \qquad \forall x \in \mathbb{R}.$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}\int_{-\infty}^{\infty} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}, \qquad \forall y \in \mathbb{R}.$$

That is, if we only keep track of the x-coordinate of the random dart, then this x-coordinate is a standard Gaussian itself. And if we only keep track of the y-coordinate of the random dart, then this y-coordinate is also a standard Gaussian.

Example 1.50 (**Buffon's Needle**)**.** Suppose a needle of length $\ell > 0$ is kept parallel to the ground. The needle is dropped onto the ground with a random position and orientation. The ground has a grid of equally spaced horizontal lines, where the distance between two adjacent lines is $d > 0$. Suppose $\ell < d$. What is the probability that the needle touches one of the lines? (Since $\ell < d$, the needle can touch at most one line.)

Let x be the distance of the midpoint of the needle from the closest line. Let θ be the acute angle formed by the needle and any horizontal line. The tip of the needle exactly touches the line when $\sin\theta = x/(\ell/2) = 2x/\ell$. So, any part of the needle touches some line if and only if $x \le (\ell/2)\sin\theta$. Since the needle has a uniformly random position and orientation, we model X, Θ as random variables with joint distribution uniform on $[0, d/2] \times [0, \pi/2]$. So,

$$f_{X,\Theta}(x,\theta) = \begin{cases} \frac{4}{\pi d}, & x \in [0,d/2] \text{ and } \theta \in [0,\pi/2] \\ 0, & \text{otherwise.} \end{cases}$$

(Note that $\iint_{\mathbb{R}^2} f_{X,\Theta}(x,\theta)dxd\theta = 1$.) And the probability that the needle touches one of the lines is

$$\begin{aligned} \iint_{0\le x\le(\ell/2)\sin\theta} f_{X,\Theta}(x,\theta)dxd\theta &= \int_{\theta=0}^{\theta=\pi/2}\int_{x=0}^{x=(\ell/2)\sin\theta} \frac{4}{\pi d}dxd\theta \\ &= \frac{2\ell}{\pi d}\int_{\theta=0}^{\theta=\pi/2}\sin\theta d\theta = \frac{2\ell}{\pi d}[-\cos\theta]_{\theta=0}^{\theta=\pi/2} = \frac{2\ell}{\pi d}. \end{aligned}$$

Note that $x \le \ell/2 < d/2$ always, so the set $0 \le x \le (\ell/2)\sin\theta$ is still contained in the set $x \in [0, d/2]$.

In particular, when $\ell = d$, the probability is $2/\pi$.

Definition 1.51. Let X, Y be random variables with joint PDF $f_{X,Y}$. Let $g\colon \mathbb{R}^2 \to \mathbb{R}$. Then

$$\mathbf{E}g(X,Y) = \iint_{\mathbb{R}^2} g(x,y)f_{X,Y}(x,y)dxdy.$$

In particular,

$$\mathbf{E}(XY) = \iint_{\mathbb{R}^2} xyf_{X,Y}(x,y)dxdy.$$

Exercise 1.52. Let X, Y be random variables with joint PDF $f_{X,Y}$. Let $a, b \in \mathbb{R}$. Using Definition 1.51, show that $\mathbf{E}(aX + bY) = a\mathbf{E}X + b\mathbf{E}Y$.

Definition 1.53 (**Joint Density Function**)**.** We say that random variables $X_1, \dots, X_n$ have **joint density function** $f\colon \mathbb{R}^n \to [0,\infty)$ if $\int_{\mathbb{R}^n} f(x)dx = 1$, and if

$$\mathbf{P}((X_1,\dots,X_n) \in A) = \int_A f(x)dx, \qquad \forall\, A \subseteq \mathbb{R}^n.$$

We define the **marginal density** $f_1\colon \mathbb{R} \to [0,\infty)$ of X_1 so that

$$f_1(x_1) = \int_{\mathbb{R}^{n-1}} f(x_1,\dots,x_n)dx_2\cdots dx_n, \qquad \forall\, x_1 \in \mathbb{R}.$$

Similarly, we can define the marginal density $f_{12}\colon \mathbb{R}^2 \to [0,\infty)$ of X_1, X_2 so that

$$f_{12}(x_1,x_2) = \int_{\mathbb{R}^{n-2}} f(x_1,\dots,x_n)dx_3\cdots dx_n, \qquad \forall\, x_1, x_2 \in \mathbb{R}.$$

And so on.

Exercise 1.54. Let X_1, Y_1 be random variables with joint PDF f_{X_1,Y_1}. Let X_2, Y_2 be random variables with joint PDF f_{X_2,Y_2}. Let $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ and let $S\colon \mathbb{R}^2 \to \mathbb{R}^2$ so that $ST(x,y) = (x,y)$ and $TS(x,y) = (x,y)$ for every $(x,y) \in \mathbb{R}^2$. Let $J(x,y)$ denote the determinant of the Jacobian of S at (x,y). Assume that $(X_2, Y_2) = T(X_1, Y_1)$. Using the change of variables formula from multivariable calculus, show that

$$f_{X_2,Y_2}(x,y) = f_{X_1,Y_1}(S(x,y))\,|J(x,y)|\,.$$

We defined independence of random variables in Definition 1.17. Below is an equivalent definition (the equivalence is beyond the scope of this course).

Definition 1.55 (**Independence of Random Variables**)**.** Let $X_1, \ldots, X_n$ be random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. We say that $X_1, \ldots, X_n$ are **independent** if

$$\mathbf{P}(X_1 \le x_1, \ldots, X_n \le x_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \le x_i), \qquad \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

Exercise 1.56. Let $X_1, \ldots, X_n$ be discrete random variables. Assume that

$$\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbf{P}(X_i = x_i), \qquad \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

Show that $X_1, \ldots, X_n$ are independent.

Exercise 1.57. Let $X_1, \ldots, X_n$ be continuous random variables with joint PDF $f\colon \mathbb{R}^n \to [0, \infty)$. Assume that

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \qquad \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

Show that $X_1, \ldots, X_n$ are independent.

Exercise 1.58. Let $X_1, \ldots, X_n\colon \Omega \to \mathbb{R}$ be uncorrelated random variables with $\mathbf{E}X_i^2 < \infty$ for any $1 \le i \le n$. Show that

$$\operatorname{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \operatorname{var}(X_i)$$

Proposition 1.59. *Let $X_1, \ldots, X_n$ be random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Assume that $X_1, \ldots, X_n$ are pairwise independent. That is, X_i and X_j are independent whenever $i, j \in \{1, \ldots, n\}$ with $i \ne j$. Then*

$$\operatorname{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \operatorname{var}(X_i).$$

Proposition 1.60. *Let $X_1, \ldots, X_n$ be independent random variables. Then*

$$\mathbf{E}(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} \mathbf{E}(X_i).$$

Proposition 1.61. *Let $0 = n_0 < n_1 < n_2 < \ldots < n_k = n$ be integers. Let $X_1, \ldots, X_n$ be independent random variables. For any $1 \leq i \leq k$, let $g_i \colon \mathbb{R}^{n_i - n_{i-1}} \to \mathbb{R}$. Then the random variables $g_1(X_1, \ldots, X_{n_1})$, $g_2(X_{n_1+1}, \ldots, X_{n_2}), \ldots, g_k(X_{n_{k-1}+1}, \ldots, X_{n_k})$ are independent. Consequently,*

$$\mathbf{E}(\prod_{i=1}^{k} g_i(X_{n_{i-1}+1}, \ldots, X_{n_i})) = \prod_{i=1}^{k} \mathbf{E} g_i(X_{n_{i-1}+1}, \ldots, X_{n_i}).$$

Definition 1.62 (**Covariance**)**.** Let X and Y be random variables with finite variances. We define the **covariance** of X and Y, denoted $\mathrm{cov}(X, Y)$, by

$$\mathrm{cov}(X, Y) = \mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))).$$

Remark 1.63. By the Cauchy-Schwarz inequality (see Theorem 1.99), we have

$$|\mathrm{cov}(X, Y)| \leq (\mathbf{E}(X - \mathbf{E}X)^2)^{1/2} (\mathbf{E}(Y - \mathbf{E}Y)^2)^{1/2}.$$

So, the covariance is well defined if X, Y both have finite variance. Note that

$$\mathrm{cov}(X, X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \mathrm{var}(X).$$

The covariance of X and Y is meant to measure whether or not X and Y are related somehow. The covariance of two random variables can be any real number. In order to more accurately measure how two random variables are "related" to each other, it is natural to divide the covariance by the product of the standard deviations, i.e. the right side of Remark 1.63.

In linear algebraic terms, if we think of the random variables $X - \mathbf{E}X$ and $Y - \mathbf{E}Y$ as vectors with the inner product $\langle X - \mathbf{E}X, Y - \mathbf{E}Y \rangle := \mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)]$ and norm $\|(X - \mathbf{E}X)\| := \langle X - \mathbf{E}X, X - \mathbf{E}X \rangle^{1/2}$, then the covariance is the cosine of the angle between the unit vectors $\frac{X - \mathbf{E}X}{\|X - \mathbf{E}X\|}$ and $\frac{Y - \mathbf{E}Y}{\|Y - \mathbf{E}Y\|}$.

Definition 1.64 (**Correlation**)**.** Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X and Y be discrete random variables on Ω taking a finite number of values. We define the **correlation** of X and Y to be

$$\frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}.$$

From Remark 1.63, the correlation of X and Y is a real number in the interval $[-1, 1]$. If the correlation is 1 or -1, then $X - \mathbf{E}X$ is a constant multiple of $Y - \mathbf{E}Y$ with probability 1, by the known equality case of the Cauchy-Schwarz inequality (see Theorem 1.99). By contrast, correlation zero is analogous to X and Y being independent. However, correlation zero does not necessarily imply that X and Y are independent. Other correlation values can be thought of as an interpolations between these extreme cases.

Exercise 1.65. Let $X_1, \ldots, X_n$ be random variables. Then

$$\mathrm{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \mathrm{cov}(X_i, X_j).$$

1.5. **Conditional Probability and Conditional Expectation.** In elementary probability theory, conditional probability and conditional expectation allow a rigorous notion for incorporating previously unknown information into a probability law.

Definition 1.66. If A, B are events and if $\mathbf{P}(B) > 0$, we define the **conditional probability of A given B**, denoted $\mathbf{P}(A|B)$, to be

$$\mathbf{P}(A|B) := \mathbf{P}(A \cap B)/\mathbf{P}(B).$$

For example, if $\mathbf{P}$ is uniform on the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$, and if $B = \{2, 4, 6\}$, then $\mathbf{P}(\{1\}|B) = 0$ and $\mathbf{P}(\{2\}|B) = 1/3$.

Let $X \colon \Omega \to [-\infty, \infty]$ be a random variable with $\mathbf{E}\,|X| < \infty$. Note that, if B is fixed, then the function $A \mapsto \mathbf{P}(A|B)$ is itself a probability law on Ω, so we can e.g. define the **conditional expectation** of a random variable X given B, denoted $\mathbf{E}(X|B)$, to be the usual expectation of X with respect to the probability law $\mathbf{P}(\cdot|B)$.

$$\mathbf{E}(X|B) := \mathbf{E}(X1_B)/\mathbf{P}(B).$$

In case $X \geq 0$, we have the equivalent definition $\mathbf{E}(X|B) = \int_0^\infty \mathbf{P}(X > t|B)dt$.

If Z is a discrete random variable, i.e. if Z takes at most countably many values, and if $\mathbf{P}(Z = z) > 0$ for some $z \in \mathbb{R}$, we let $B := \{Z = z\}$ in the above definition to define $\mathbf{E}(X|Z = z)$. By splitting the sample space Ω into countably many disjoint sets $B_1, B_2, \ldots$ such that $\cup_{n=1}^\infty B_n = \Omega$ and $\mathbf{P}(B_n) > 0$ for all $n \geq 1$, we can write

$$\mathbf{P}(A) = \sum_{n=1}^\infty \mathbf{P}(A \cap B_n) = \sum_{n=1}^\infty \mathbf{P}(A|B_n)\mathbf{P}(B_n).$$

$$\mathbf{E}X = \sum_{n=1}^\infty \mathbf{E}(X1_{B_n}) = \sum_{n=1}^\infty \mathbf{E}(X|B_n)\mathbf{P}(B_n). \tag{1}$$

By breaking up expected values or probabilities into pieces in this way, sometimes the quantities on the right side are easier to compute, allowing computation of the left side.

There is a way to condition on events with probability zero, but we will not do so here.

Proposition 1.67. *Let B be a fixed subset of some sample space Ω. Let $\mathbf{P}$ be a probability law on Ω. Assume that $\mathbf{P}(B) > 0$. Given any subset A in Ω, define $\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$ as above. Then $\mathbf{P}(A|B)$ is itself a probability law on Ω.*

Proof. We first verify Axiom (i). Let $A \subseteq \Omega$. Since Axiom (i) holds for $\mathbf{P}$ by assumption, we have $\mathbf{P}(A \cap B) \geq 0$. Therefore, $\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B) \geq 0$.

We now verify Axiom (iii). Note that $\mathbf{P}(\Omega|B) = \mathbf{P}(\Omega \cap B)/\mathbf{P}(B) = \mathbf{P}(B \cap B)/\mathbf{P}(B) = \mathbf{P}(B)/\mathbf{P}(B) = 1$.

We now verify Axiom (ii). Let $A, C \subseteq \Omega$ with $A \cap C = \emptyset$. Since A and C are disjoint, we know that $A \cap B$ and $C \cap B$ are disjoint. So, we can apply Axiom (ii) for $\mathbf{P}$ to the sets $A \cap B$ and $C \cap B$. So,

$$\begin{aligned}\mathbf{P}(A \cup C|B)\mathbf{P}(B) &= \mathbf{P}((A \cup C) \cap B) = \mathbf{P}((A \cap B) \cap (C \cap B)), \quad \text{by Proposition 1.6(ii)} \\ &= \mathbf{P}(A \cap B) + \mathbf{P}(C \cap B) = \mathbf{P}(A|B)\mathbf{P}(B) + \mathbf{P}(C|B)\mathbf{P}(B).\end{aligned}$$

Dividing both sides by $\mathbf{P}(B)$ implies that Axiom (ii) holds for two sets. To verify that additivity holds for a countable number of sets, let $A_1, A_2, \ldots$ be subsets of Ω such that

$A_i \cap A_j = \emptyset$ whenever i, j are positive integers with $i \neq j$. Since $A_i \cap A_j = \emptyset$ whenever $i \neq j$, we have $(A_i \cap B) \cap (A_j \cap B) = \emptyset$. So, using Exercise 1.9, and Axiom (ii) for $\mathbf{P}$,

$$\mathbf{P}(B)\mathbf{P}\left(\bigcup_{k=1}^{\infty} A_k \,\middle|\, B\right) = \mathbf{P}\left(\left(\bigcup_{k=1}^{\infty} A_k\right) \cap B\right) = \mathbf{P}\left(\bigcup_{k=1}^{\infty} (A_k \cap B)\right), \quad \text{by Exercise 1.9}$$
$$= \sum_{k=1}^{\infty} \mathbf{P}(A_k \cap B) = \mathbf{P}(B) \sum_{k=1}^{\infty} \mathbf{P}(A_k|B)$$

So, Axiom (ii) holds. In conclusion, $\mathbf{P}(A|B)$ is a probability law on Ω. □

Remark 1.68. Proposition 1.67 implies that facts from Proposition 1.13 apply also to conditional probabilities. For example, using the notation of Proposition 1.67, we have $\mathbf{P}(A \cup C|B) \leq \mathbf{P}(A|B) + \mathbf{P}(C|B)$.

Example 1.69 (Medical Testing). Suppose a test for a disease is 99% accurate. That is, if you have the disease, the test will be positive with 99% probability. And if you do not have the disease, the test will be negative with 99% probability. Suppose also the disease is fairly rare, so that roughly 1 in 10,000 people have the disease. If you test positive for the disease, with what probability do you actually have the disease?

The answer is unfortunately around 1/100. To see this, let's consider the probabilities. Let B be the event that you test positive for the disease. Let A be the event that you actually have the disease. We want to compute $\mathbf{P}(A|B)$. We have

$$\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B) = (\mathbf{P}(A)/\mathbf{P}(B))\mathbf{P}(A \cap B)/\mathbf{P}(A) = (\mathbf{P}(A)/\mathbf{P}(B))\mathbf{P}(B|A).$$

We are given that $\mathbf{P}(A) = 10^{-4}$, $P(B|A) = .99$ and $\mathbf{P}(B|A^c) = .01$. To compute $\mathbf{P}(B)$, we write $B = (B \cap A) \cup (B \cap A^c)$, so that

$$\mathbf{P}(B) = \mathbf{P}(B \cap A) + \mathbf{P}(B \cap A^c) = \mathbf{P}(B|A)\mathbf{P}(A) + \mathbf{P}(B|A^c)\mathbf{P}(A^c)$$
$$= .99(10^{-4}) + .01(1 - \mathbf{P}(A)) = .99(10^{-4}) + .01(1 - 10^{-4}) \approx 10^{-2}.$$

In conclusion,

$$\mathbf{P}(A|B) = \frac{10^{-4}}{\mathbf{P}(B)}(.99) \approx 10^{-4}10^{2} = 10^{-2}.$$

So, even though the test is fairly accurate from a certain perspective, a positive test result does not say very much.

Many people find this result counterintuitive, though the following reasoning can help to explain the result. Suppose we have a population of 10,000 people. Then roughly 1 person in the population has the disease. Suppose everyone is given the test. Since 9,999 people are healthy and the test is 99% accurate, around 100 healthy people will test positive for the disease. Meanwhile, the 1 sick person will most likely test positive for the disease. So, out of around 101 people testing positive for the disease, only 1 of them actually has the disease. So, $\mathbf{P}(A|B)$ is roughly $1/101 \approx 10^{-2}$.

Definition 1.70 (Conditioning a Continuous Random Variable on a Set). Let X be a continuous random variable on a sample space Ω. Let $A \subseteq \Omega$ with $\mathbf{P}(A) > 0$. The **conditional PDF** $f_{X|A}$ of X given A is defined to be the function $f_{X|A}$ satisfying

$$\mathbf{P}(X \in B\,|A) = \int_B f_{X|A}(x)dx, \qquad \forall\, B \subseteq \mathbb{R}.$$

Example 1.71. Suppose $A' \subseteq \mathbb{R}$ and we condition on X satisfying $X \in A'$. That is, A is the event $A = \{X \in A'\}$. Then, using Definition 1.66,

$$\mathbf{P}(X \in B \,|A) = \mathbf{P}(X \in B \,|X \in A') = \frac{\mathbf{P}(X \in B, X \in A')}{\mathbf{P}(X \in A')} = \frac{\int_{B \cap A'} f_X(x)dx}{\mathbf{P}(X \in A')}.$$

So, using Definition 1.70, in this case we have

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{\mathbf{P}(X \in A')}, & x \in A' \\ 0, & \text{otherwise.} \end{cases}$$

Example 1.72. Suppose you go to the bus stop, and the time T between successive arrivals of the bus is an exponential random variable with parameter $\lambda > 0$. Let $t > 0$. Suppose you go to the bus stop and someone says the last bus came t minutes ago. Let A be the event that $T > t$. That is, we will take it as given that $T > t$, i.e. that up to time t, the bus has not yet arrived. Let X be the time you need to wait until the next bus arrives. Let $x > 0$. Using Definition 1.66 and Example 1.25,

$$\begin{aligned} \mathbf{P}(X > x|A) = \mathbf{P}(T > t + x|T > t) &= \frac{\mathbf{P}(T > t + x, T > t)}{\mathbf{P}(T > t)} = \frac{\mathbf{P}(T > t + x)}{\mathbf{P}(T > t)} \\ &= \frac{\lambda \int_{t+x}^{\infty} e^{-\lambda s} ds}{\lambda \int_t^{\infty} e^{-\lambda s} ds} = \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x} = \lambda \int_x^{\infty} e^{-\lambda s} ds. \end{aligned}$$

From Definition 1.70, $\mathbf{P}(X > x|A) = \int_x^{\infty} f_{X|A}(x)dx$. That is, $f_{X|A}(x) = \lambda e^{-\lambda x}$. That is, $X|A$ is also an exponential random variable with parameter λ. That is, even though we know the bus has not arrived for t minutes, this does not at all affect our prediction for the arrival of the next bus.

This property is called the **memoryless** property of the exponential random variable.

Exercise 1.73. Suppose you go to the bus stop, and the time T between successive arrivals of the bus is anything between 0 and 30 minutes, with all arrival times being equally likely.

Suppose you get to the bus stop, and the bus just leaves as you arrive. How long should you expect to wait for the next bus? What is the probability that you will have to wait at least 15 minutes for the next bus to arrive?

On a different day, suppose you go to the bus stop and someone says the last bus came 10 minutes ago. How long should you expect to wait for the next bus? What is the probability that you will have to wait at least 10 minutes for the next bus to arrive?

Exercise 1.74. Let $A_1, A_2, \ldots$ be disjoint events such that $\mathbf{P}(A_i) = 2^{-i}$ for each $i \geq 1$. Assume $\cup_{i=1}^{\infty} A_i = \Omega$. Let X be a random variable such that $\mathbf{E}(X|A_i) = (-1)^{i+1}$ for each $i \geq 1$. Compute $\mathbf{E}X$.

Definition 1.75 (Conditioning one Random Variable on Another). Let X and Y be continuous random variables with joint PDF $f_{X,Y}$. Fix some $y \in \mathbb{R}$ with $f_Y(y) > 0$. For any $x \in \mathbb{R}$, define the **conditional PDF** of X, given that $Y = y$ by

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \qquad \forall\, x \in \mathbb{R}.$$

We also define the **conditional expectation** of X given $Y = y$ by

$$\mathbf{E}(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx.$$

From Definition 1.48, note that $\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = 1$. So, $f_{X|Y}(x|y)$ is a probability distribution function.

Example 1.76. We continue the dart board example from Example 1.49. Suppose X and Y have a joint PDF so that

$$\mathbf{P}((X,Y) \in A) = \frac{1}{2\pi} \iint_A e^{-(x^2+y^2)/2} dxdy, \qquad \forall\, A \subseteq \mathbb{R}^2.$$

We verified the marginals are both standard Gaussians:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad \forall\, x \in \mathbb{R}, \qquad f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \qquad \forall\, y \in \mathbb{R}.$$

So, in this particular example, we have

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{1}{2\pi} e^{-(x^2+y^2)/2}}{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

That is, in this particular example, conditioning X on Y does not at all change X.

Example 1.77. Suppose X and Y have a joint PDF given by $f_{X,Y}(x,y) = \frac{1}{\pi}$ if $x^2 + y^2 \leq 1$, and $f_{X,Y}(x,y) = 0$ otherwise. Let's compute the marginals first, and then determine the conditional PDFs. Let $x, y \in \mathbb{R}$ with $x^2 + y^2 \leq 1$. Using Definition 1.48,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{y=-\sqrt{1-x^2}}^{y=\sqrt{1-x^2}} \frac{1}{\pi} dy = \frac{2\sqrt{1-x^2}}{\pi}.$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_{x=-\sqrt{1-y^2}}^{x=\sqrt{1-y^2}} \frac{1}{\pi} dx = \frac{2\sqrt{1-y^2}}{\pi}.$$

So, if $x^2 + y^2 \leq 1$, then

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1/\pi}{2\sqrt{1-y^2}/\pi} = \frac{1}{2\sqrt{1-y^2}}.$$

Similarly,

$$f_{Y|X}(y|x) = \frac{1}{2\sqrt{1-x^2}}.$$

That is, in this particular example, conditioning X on Y can drastically change X. For example, X conditioned on $Y = 0$, and X conditioned on $Y = 1/2$ have very different PDFs.

The following Theorem is a version of (1) for continuous random variables.

Theorem 1.78 (**Total Expectation Theorem**). *Let X, Y be continuous random variables. Assume that $f_{X,Y} \colon \mathbb{R}^2 \to \mathbb{R}$ is a continuous function. Then*

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} \mathbf{E}(X|Y = y) f_Y(y) dy.$$

Proof. Using Definition 1.75 and then Definition 1.48,

$$\begin{aligned}\int_{-\infty}^{\infty} \mathbf{E}(X|Y=y) f_Y(y) dy &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx \right) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y) dy \right) dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) dy dx \\ &= \int_{-\infty}^{\infty} x f_X(x) dx = \mathbf{E}X.\end{aligned}$$

□

In the above proof, we used the following Theorem from analysis.

Theorem 1.79 (**Fubini Theorem**). *Let $h\colon \mathbb{R}^2 \to \mathbb{R}$ be a continuous function such that $\iint_{\mathbb{R}^2} |h(x,y)|\, dxdy < \infty$. Then*

$$\iint_{\mathbb{R}^2} h(x,y) dxdy = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x,y) dx \right) dy = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x,y) dy \right) dx.$$

Theorem 1.80 (**Fubini Theorem for Sums**). *Let $\{a_{ij}\}_{i,j\geq 0}$ be a doubly-infinite array of nonnegative numbers. Then*

$$\sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} a_{ij} \right) = \sum_{j=0}^{\infty} \left(\sum_{i=0}^{\infty} a_{ij} \right).$$

Exercise 1.81. Find a doubly-infinite array of real numbers $\{a_{ij}\}_{i,j\geq 0}$ such that

$$\sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} a_{ij} \right) = 1 \neq 0 = \sum_{j=0}^{\infty} \left(\sum_{i=0}^{\infty} a_{ij} \right).$$

(Hint: the array can be chosen to have all entries either $-1, 0$, or 1. And most of the entries can be chosen to be 0.)

Exercise 1.82. Let X, Y be random variables. For any $y \in \mathbb{R}$, assume that $\mathbf{E}(X|Y=y) = e^{-|y|}$. Also, assume that Y has an exponential distribution with parameter $\lambda = 2$. Compute $\mathbf{E}X$.

1.6. **Functions of Random Variables.**

Proposition 1.83. *Let X be a continuous random variable with density function $f_X\colon \mathbb{R} \to [0,\infty)$. Let $g\colon \mathbb{R} \to \mathbb{R}$ be continuous. Let $Y := g(X)$. Assume that f_X is a continuous function. Then for any $y \in \mathbb{R}$,*

$$f_Y(y) = \frac{d}{dy} \int_{\{x\in\mathbb{R}\colon\, g(x)\leq y\}} f_X(x) dx.$$

Proof. Let $A \subseteq \mathbb{R}$. Recall that f_X is defined so that

$$\mathbf{P}(X \in A) = \int_A f_X(x) dx.$$

So, if we let $y \in \mathbb{R}$ and if we define $A := \{x \in \mathbb{R}\colon g(x) \leq y\}$, we have

$$F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(g(X) \leq y) = \mathbf{P}(X \in A) = \int_A f_X(x) dx = \int_{\{x\in\mathbb{R}\colon\, g(x)\leq y\}} f_X(x) dx.$$

So, if F_Y is differentiable, $\frac{d}{dy}F_Y(y) = f_Y(y)$ for all $y \in \mathbb{R}$, completing the proof by the Fundamental Theorem of Calculus, Theorem 1.42. $\square$

Example 1.84. Let X be a uniformly distributed random variable on $[-1, 1]$, and let $g\colon \mathbb{R} \to \mathbb{R}$ so that $g(x) = x^3$ for any $x \in \mathbb{R}$. Let $Y := g(X)$. Then for any $y \in \mathbb{R}$,

$$f_Y(y) = \frac{d}{dy}\int_{\{x\in\mathbb{R}\colon\, g(x)\le y\}} f_X(x)dx = \frac{d}{dy}\int_{\{x\in[-1,1]\colon\, x^3\le y\}} \frac{1}{2}dx.$$

If $y < -1$ the integral is zero. If $y > 1$, the integral is 1. And if $y \in [-1, 1]$, we have

$$f_Y(y) = \frac{d}{dy}\frac{1}{2}\int_{x=-1}^{x=y^{1/3}} dx = \frac{1}{2}\frac{d}{dy}[y^{1/3} + 1] = \frac{1}{6}y^{-2/3}.$$

And if $y \notin [-1, 1]$, we have $f_Y(y) = 0$.

Definition 1.85 (**Monotonic Function**)**.** Let $I, J \subseteq \mathbb{R}$ be open intervals. Let $g\colon I \to J$. We say that g is **strictly increasing** if, for any $x, y \in I$ with $x > y$, we have $g(x) > g(y)$. We say that g is **strictly decreasing** if, for any $x, y \in I$ with $x > y$, we have $g(x) < g(y)$. We say that g is **strictly monotonic** if g is either strictly increasing or strictly decreasing.

Remark 1.86 (**Monotonic Functions are Invertible**)**.** Let $I, J \subseteq \mathbb{R}$ be open intervals. Let $g\colon I \to J$ be a monotonic function with range J. As we recall from calculus, g has an inverse. That is, there exists a monotonic function $h\colon J \to I$ such that $g(h(x)) = x$ for every $x \in J$ and $h(g(x)) = x$ for every $x \in I$. Also, as we recall from calculus, if g is differentiable with $g'(x) \neq 0$ for all $x \in I$, then h is differentiable, and by differentiating the identity $h(g(x)) = x$ and applying the chain rule, we get

$$\frac{d}{dx}h(g(x)) = \frac{1}{g'(x)}, \qquad \forall\, x \in I.$$

Or, written another way (defining $y := g(x)$, so that $x = h(y)$),

$$h'(y) = \frac{1}{g'(h(y))}, \qquad \forall\, y \in J.$$

If we graph g and h, then h is obtained by reflecting g across the line $\{(x, y) \in \mathbb{R}^2\colon x = y\}$. Similarly, g is obtained by reflecting h across the line $\{(x, y) \in \mathbb{R}^2\colon x = y\}$.

Proposition 1.87. *Let X be a continuous random variable such that F_X is differentiable. Let $I, J \subseteq \mathbb{R}$ be open intervals. Let $g\colon I \to J$ be a monotonic, differentiable function with range J. Assume that $g'(x) \neq 0$ for every $x \in I$. Let $Y := g(X)$. Let $h\colon J \to I$ be the inverse of g. Then for any $y \in J$,*

$$f_Y(y) = f_X(h(y)) \cdot \left|\frac{d}{dy}h(y)\right| = f_X(h(y)) \cdot \frac{1}{|g'(h(y))|}.$$

Proof. Let $y \in J$. First, assume g is strictly increasing. Then

$$F_Y(y) = \mathbf{P}(Y \le y) = \mathbf{P}(g(X) \le y) = \mathbf{P}(X \le h(y)) = F_X(h(y)).$$

Since F_X and h are differentiable, the Chain Rule then proves the first equality, using also the Fundamental Theorem of Calculus, Theorem 1.42.. The second equality follows from

Remark 1.86, where we noted that

$$\frac{d}{dy}h(y) = \frac{1}{g'(h(y))}, \qquad \forall y \in J.$$

□

Exercise 1.88. Let X be a uniformly distributed random variable on $[0, 1]$. Find the PDF of $-\log(X)$.

Exercise 1.89. Let X be a standard normal random variable. Find the PDF of e^X.

1.7. **Inequalities.**

Exercise 1.90. Let $\phi\colon \mathbb{R} \to \mathbb{R}$. We say that ϕ is **convex** if, for any $x, y \in \mathbb{R}$ and for any $t \in [0, 1]$, we have

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y).$$

Let $\phi\colon \mathbb{R} \to \mathbb{R}$. Show that ϕ is convex if and only if: for any $y \in \mathbb{R}$, there exists a constant a and there exists a function $L\colon \mathbb{R} \to \mathbb{R}$ defined by $L(x) = a(x-y) + \phi(y)$, $x \in \mathbb{R}$, such that $L(y) = \phi(y)$ and such that $L(x) \leq \phi(x)$ for all $x \in \mathbb{R}$. (In the case that ϕ is differentiable, the latter condition says that ϕ lies above all of its tangent lines.)

(Hint: Suppose ϕ is convex. If x is fixed and y varies, show that $\frac{\phi(y)-\phi(x)}{y-x}$ increases as y increases. Draw a picture. What slope a should L have at x?)

Exercise 1.91 (**Jensen's Inequality**)**.** Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be convex. Assume that $\mathbf{E}\,|X| < \infty$ and $\mathbf{E}\,|\phi(X)| < \infty$. Then

$$\phi(\mathbf{E}X) \leq \mathbf{E}\phi(X).$$

(Hint: use Exercise 1.90 with $y := \mathbf{E}X$.) Deduce the **triangle inequality**:

$$|\mathbf{E}X| \leq \mathbf{E}\,|X|\,.$$

Exercise 1.92 (**Markov's Inequality**)**.** Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable. Then

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|}{t}, \qquad \forall\, t > 0.$$

(Hint: multiply both sides by t and use monotonicity of $\mathbf{E}$.)

Corollary 1.93. *If n is a positive integer, then*

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|^n}{t^n}, \qquad \forall\, t > 0.$$

Proof. From Markov's Inequality, Exercise 1.92,

$$\mathbf{P}(|X| \geq t) = \mathbf{P}(|X|^n \geq t^n) \leq \frac{\mathbf{E}\,|X|^n}{t^n}, \qquad \forall\, t > 0.$$

□

We refer to $\mathbf{E}\,|X|^n$ as the n^{th} **moment** of X.

Definition 1.94 (**Variance**)**.** Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable with $\mathbf{E}\,|X| < \infty$ and $\mathbf{E}X^2 < \infty$. We define the **variance** of X, denoted $\mathrm{var}(X)$, to be

$$\mathrm{var}(X) := \mathbf{E}(X - \mathbf{E}X)^2 = \mathbf{E}X^2 - (\mathbf{E}X)^2.$$

Remark 1.95. By Jensen's Inequality, if $\mathbf{E}X^2 < \infty$, then $\mathbf{E}|X| < \infty$, so $\mathbf{E}X \in \mathbb{R}$.

Exercise 1.96. Let $a, b \in \mathbb{R}$ and let $X\colon \Omega \to [-\infty, \infty]$ be a random variable with $\mathbf{E}X^2 < \infty$. Show that

$$\mathrm{var}(aX + b) = a^2 \mathrm{var}(X).$$

Then, let X be a standard Gaussian. Show that $\mathbf{E}X = 0$ and $\mathrm{var}(X) = 1$.

Finally, show that the quantity $\mathbf{E}(X - t)^2$ is minimized for $t \in \mathbb{R}$ uniquely when $t = \mathbf{E}X$.

Replacing X by $X - \mathbf{E}X$ and taking $n = 2$ in Corollary 1.93 gives:

Corollary 1.97 (Chebyshev's Inequality). *Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable with $\mathbf{E}X^2 < \infty$. Then*

$$\mathbf{P}(|X - \mathbf{E}X| \geq t) \leq \frac{\mathrm{var}(X)}{t^2}, \qquad \forall\, t > 0.$$

(By Exercise 1.91, $\mathbf{E}X \in \mathbb{R}$.)

Corollary 1.93 shows that, if large moments of X are finite, then $\mathbf{P}(X > t)$ decays rapidly. Sometimes, we can even get exponential decay on $\mathbf{P}(X > t)$, if we make the rather strong assumption that $\mathbf{E}e^{rX}$ is finite for some $r > 0$. Note that, by the power series expansion of the exponential, $\mathbf{E}e^{rX} < \infty$ assumes that an infinite sum of the moments of X is finite.

Exercise 1.98 (The Chernoff Bound). Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable. Show that, for any $r, t > 0$,

$$\mathbf{P}(X > t) \leq e^{-rt}\mathbf{E}e^{rX}.$$

If $1 \leq p < \infty$, and if $X\colon \Omega \to [-\infty, \infty]$ is a random variable, denote the **L_p-norm** of X as $\|X\|_p := (\mathbf{E}|X|^p)^{1/p}$ and denote the **L_∞-norm** of X as $\|X\|_\infty := \inf\{c > 0\colon \mathbf{P}(|X| \leq c) = 1\}$.

Theorem 1.99 (Hölder's Inequality). *Let $X, Y\colon \Omega \to \mathbb{R}$ be random variables. Let $1 \leq p \leq \infty$, and let q be dual to p (so $1/p + 1/q = 1$). Then*

$$\mathbf{E}|XY| \leq \|X\|_p \|Y\|_q.$$

*This inequality is an equality only if X is a constant multiple of Y with probability 1. The case $p = q = 2$ recovers the **Cauchy-Schwarz** inequality:*

$$\mathbf{E}|XY| \leq (\mathbf{E}X^2)^{1/2}(\mathbf{E}Y^2)^{1/2}.$$

Proof. By scaling, we may assume $\|X\|_p = \|Y\|_q = 1$ (zeros and infinities being trivial). Also, the case $p = 1, q = \infty$ follows from the triangle inequality, so we assume $1 < p < \infty$. From concavity of the log function, we have the pointwise inequality

$$|X(\omega)Y(\omega)| = (|X(\omega)|^p)^{1/p}(|Y(\omega)|^q)^{1/q} \leq \frac{1}{p}|X(\omega)|^p + \frac{1}{q}|Y(\omega)|^q, \qquad \forall\, \omega \in \Omega$$

which upon integration gives the result. If this inequality is an equality with probability one, then the strict concavity of the log function implies that $\mathbf{P}(X = Y) = 1$. □

Theorem 1.100 (Triangle Inequality). *Let $X, Y\colon \Omega \to \mathbb{R}$ be random variables. Let $1 \leq p \leq \infty$. Then*

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p, \quad 1 \leq p \leq \infty$$

Proof. The case $p = \infty$ follows from the scalar triangle inequality, so assume $1 \leq p < \infty$. By scaling, we may assume $\|X\|_p = 1 - t$, $\|Y\|_p = t$, for some $t \in (0, 1)$ (zeros and infinities being trivial). Define $V := X/(1-t)$, $W := Y/t$. Then by convexity of $x \mapsto |x|^p$ on $\mathbb{R}$,

$$|(1-t)V(\omega) + tW(\omega)|^p \leq (1-t)\,|V(\omega)|^p + t\,|W(\omega)|^p\,, \qquad \forall\, \omega \in \Omega$$

which upon integration completes the proof. □

Exercise 1.101. Let $X, Y \colon \Omega \to \mathbb{R}$ be random variables. Let $0 < p < 1$ and let $\|X\|_p := (\mathbf{E}\,|X|^p)^{1/p}$. Show that there exists $c(p) > 0$ such that $\|X + Y\|_p \leq c(p)(\|X\|_p + \|Y\|_p)$. In particular, it suffices to choose $c(p) = 2^{1/p}$. (Hint: a pointwise inequality should imply that $\|X + Y\|_p^p \leq \|X\|_p^p + \|Y\|_p^p$.)

Exercise 1.102 (MAX-CUT). The probabilistic method is a very useful way to prove the existence of something satisfying some properties. This method is based upon the following elementary statement: If $\alpha \in \mathbb{R}$ and if a random variable $X \colon \Omega \to \mathbb{R}$ satisfies $\mathbf{E}X \geq \alpha$, then there exists some $\omega \in \Omega$ such that $X(\omega) \geq \alpha$. We will demonstrate this principle in this exercise.

Let $G = (V, E)$ be an undirected graph on the vertices $V = \{1, \ldots, n\}$ so that the edge set E is a subset of unordered pairs $\{i, j\}$ such that $i, j \in V$ and $i \neq j$. Let $S \subseteq V$ and denote $S^c := V \setminus S$. We refer to (S, S^c) as a cut of the graph G. The goal of the MAX-CUT problem is to maximize the number of edges going between S and S^c over all cuts of the graph G.

Prove that there exists a cut (S, S^c) of the graph such that the number of edges going between S and S^c is at least $|E|/2$. (Hint: define a random $S \subseteq V$ such that, for every $i \in V$, $\mathbf{P}(i \in S) = 1/2$, and the events $1 \in S, 2 \in S, \ldots, n \in S$ are all independent. If $\{i, j\} \in E$, show that $\mathbf{P}(i \in S, j \notin S) = 1/4$. So, what is the expected number of edges $\{i, j\} \in E$ such that $i \in S$ and $j \notin S$?)

1.8. **Independent Sums and Convolution.** Let X, Y be independent random variables. From Proposition 1.67, the moment generating function of $X + Y$ can be easily expressed as $M_{X+Y}(t) = M_X(t)M_Y(t)$, for any t such that both quantities on the right exist. On the other hand, the CDF of $X + Y$ has a more complicated dependence on X and Y.

Example 1.103. Let X, Y be independent integer-valued random variables. Then, repeatedly using properties of probability laws, and using that X, Y are independent,

$$\begin{aligned}\mathbf{P}(X + Y = t) &= \sum_{j,k \in \mathbb{Z}\colon\, j+k=t} \mathbf{P}(X = j, Y = k) = \sum_{j \in \mathbb{Z}} \mathbf{P}(X = j, Y = t - j) \\ &= \sum_{j \in \mathbb{Z}} \mathbf{P}(X = j)\mathbf{P}(Y = t - j) = \sum_{j \in \mathbb{Z}} p_X(j)p_Y(t - j).\end{aligned}$$

Definition 1.104 (Convolution on the integers). Let $g, h \colon \mathbb{Z} \to \mathbb{R}$ be functions. The **convolution** of g and h, denoted $g * h$, is a function $g * h \colon \mathbb{Z} \to \mathbb{R}$ defined by

$$(g * h)(t) := \sum_{j \in \mathbb{Z}} g(j)h(t - j), \qquad \forall\, t \in \mathbb{Z}.$$

Example 1.105. Let $g(k) := e^{-k}$ and let $h(k) := e^{-k}$ for any nonnegative integer $k \geq 0$, and let $g(k) = h(k) = 0$ for any other integer $k < 0$. Then if $t \geq 0$ is an integer,

$$(g * h)(t) = \sum_{k \in \mathbb{Z}} g(k)h(t-k) = \sum_{k=0}^{t} e^{-k}e^{-(t-k)} = \sum_{k=0}^{t} e^{-t} = (t+1)e^{-t}.$$

And $(g * h)(t) = 0$ for any negative integer t.

A similar formula holds for continuous random variables. That is, if X, Y are two continuous random variables, then the density of $X + Y$ is the convolution of f_X and f_Y.

Definition 1.106 (Convolution on the real line). Let $g, h \colon \mathbb{R} \to \mathbb{R}$ be functions. The **convolution** of g and h, denoted $g * h$, is a function $g * h \colon \mathbb{R} \to \mathbb{R}$ defined by

$$(g * h)(t) := \int_{-\infty}^{\infty} g(x)h(t-x)dx, \qquad \forall\, t \in \mathbb{R}.$$

Proposition 1.107. *Let X, Y be two continuous independent random variables. Assume that f_Y is a continuous function. Then*

$$f_{X+Y}(t) = (f_X * f_Y)(t), \qquad \forall\, t \in \mathbb{R}.$$

Proof. Let X, Y be independent continuous random variables. Then, changing variables,

$$\mathbf{P}(X + Y \leq t) = \int_{\{(x,y) \in \mathbb{R}^2 \colon x+y \leq t\}} f_{X,Y}(x,y)dxdy = \int_{x=-\infty}^{x=\infty} \int_{y=-\infty}^{y=t-x} f_X(x)f_Y(y)dydx.$$

Then, since $\mathbf{P}(X + Y \leq t)$ is differentiable with respect to t, we have by the Fundamental Theorem of Calculus, Theorem 1.42,

$$f_{X+Y}(t) = \frac{d}{dt}\mathbf{P}(X+Y \leq t) = \int_{x=-\infty}^{x=\infty} f_X(x) \frac{d}{dt} \int_{y=-\infty}^{y=t-x} f_Y(y)dydx = \int_{x=-\infty}^{x=\infty} f_X(x)f_Y(t-x)dx.$$

□

Example 1.108. Let $g(x) = h(x) := \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for any $x \in \mathbb{R}$. Then if $t \in \mathbb{R}$, we complete the square and change variables twice to get

$$\begin{aligned}
(g * h)(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2/2} e^{-(t-x)^2/2} dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2 + xt - t^2/2} dx \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-t/2)^2 + t^2/4 - t^2/2} dx = e^{-t^2/4} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-t/2)^2} dx \\
&= e^{-t^2/4} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2} dx = e^{-t^2/4} \frac{1}{2\sqrt{\pi}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx = e^{-t^2/4} \frac{1}{2\sqrt{\pi}}.
\end{aligned}$$

And $(g * h)(t) = e^{-t^2/4} \frac{1}{2\sqrt{\pi}}$ for any $t \in \mathbb{R}$.

Alternatively, we know that if X, Y are independent standard Gaussian random variables, then $X + Y$ is a Gaussian random variable with mean zero and variance $\sigma^2 = 2$. That is, $X + Y$ has density $e^{-t^2/4} \frac{1}{2\sqrt{\pi}}$, $t \in \mathbb{R}$.

More generally, the above argument shows: if X is a Gaussian with mean $\mu_X \in \mathbb{R}$ and variance $\sigma_X^2 > 0$, if Y is a Gaussian with mean μ_Y and variance σ_Y^2, and if X, Y are independent, then $X + Y$ is a Gaussian with mean $\mu_X + \mu_X$ and variance $\sigma_X^2 + \sigma_Y^2$. By induction, this statement implies: if $X_1, \ldots, X_n$ are independent Gaussian random variables with means

$\mu_{X_1}, \ldots, \mu_{X_n} \in \mathbb{R}$ and variances $\sigma^2_{X_1}, \ldots, \sigma^2_{X_n} > 0$, then $X_1 + \cdots + X_n$ is a Gaussian with mean $\sum_{i=1}^n \mu_{X_i}$ and variance $\sum_{i=1}^n \sigma^2_{X_i}$.

Example 1.109. Let X, Y be independent standard Gaussian random variables. We will find the distribution of $X^2 + Y^2$. First, if $t > 0$, note that

$$\mathbf{P}(X^2 \leq t) = \mathbf{P}(X \leq \sqrt{t}) = \int_{-\sqrt{t}}^{\sqrt{t}} e^{-x^2/2} dx/\sqrt{2\pi}.$$

So, if $t > 0$,

$$f_{X^2}(t) = \frac{d}{dt} 2 \int_0^{\sqrt{t}} e^{-x^2/2} dx/\sqrt{2\pi} = e^{-t/2} t^{-1/2} \frac{1}{\sqrt{2\pi}}.$$

For $t < 0$, $f_{X^2}(t) = 0$. The same formula holds for Y^2. Therefore,

$$\begin{aligned} f_{X^2+Y^2}(t) = f_{X^2} * f_{Y^2}(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x/2} x^{-1/2} e^{-(t-x)/2} (t-x)^{-1/2} 1_{x>0} 1_{t-x>0} dx \\ &= \frac{1}{2\pi} \int_0^t e^{-x/2} x^{-1/2} e^{-(t-x)/2} (t-x)^{-1/2} dx = \frac{1}{2\pi} e^{-t/2} \int_0^t x^{-1/2} (t-x)^{-1/2} dx \\ &= \frac{1}{2\pi} e^{-t/2} 2 \sin^{-1}(x^{1/2} t^{-1/2}) |_{x=0}^{x=t} = \frac{1}{2} e^{-t/2}. \end{aligned}$$

Exercise 1.110 (Convolution is Associative). Let $g, h, d \colon \mathbb{R} \to \mathbb{R}$. Then for any $t \in \mathbb{R}$,

$$((g * h) * d)(t) = (g * (h * d))(t)$$

Exercise 1.111. Let X, Y, Z be independent and uniformly distributed on $[0, 1]$. Note that f_X is not a continuous function.

Using convolution, compute f_{X+Y}. Draw f_{X+Y}. Note that f_{X+Y} is a continuous function, but it is not differentiable at some points.

Using convolution, compute f_{X+Y+Z}. Draw f_{X+Y+Z}. Note that f_{X+Y+Z} is a differentiable function, but it does not have a second derivative at some points.

Make a conjecture about how many derivatives $f_{X_1+\cdots+X_n}$ has, where $X_1, \ldots, X_n$ are independent and uniformly distributed on $[0, 1]$. You do not have to prove this conjecture. The idea of this exercise is that convolution is a kind of average of functions. And the more averaging you do, the more derivatives $f_{X_1+\cdots+X_n}$ has.

Exercise 1.112. Construct two random variables X, Y such that X and Y are each uniformly distributed on $[0, 1]$, and such that $\mathbf{P}(X + Y = 1) = 1$.

Then construct two random variables W, Z such that W and Z are each uniformly distributed on $[0, 1]$, and such that $W + Z$ is uniformly distributed on $[0, 2]$.

(Hint: there is a way to do each of the above problems with about one line of work. That is, there is a way to solve each problem without working very hard.)

1.9. **Additional Comments.** The foundations of measure theory were developed in the late 1800s and early 1900s by several mathematicians. Measure theory allows the definition of a probability law. In the 1930s, Kolmogorov provided an axiomatic foundation of probability theory via measure theory, e.g. the axioms of Definition 1.11. Probability theory was often not considered a "serious" subject, perhaps due to its historical affiliation with gambling. Since the 1930s and continuing to the present, more and more subjects embrace probabilistic

and statistical thinking. Statistics began to use more probability theory in the 1800s and 1900s.

2. Limit Theorems

The Laws of Large Numbers and Central Limit Theorem provide limiting statements for sequences of random variables. The exact notions of convergence will depend on the limit theorem. The general goal is to obtain the strongest possible convergence with the weakest possible assumption. Sometimes, the convergence can be upgraded to a stronger notion, but other times this is impossible.

2.1. Modes of Convergence.

Below are a few of the most commonly encountered notions of convergence of random variables.

Definition 2.1 (**Almost Sure Convergence**)**.** We say random variables $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ converge **almost surely** (or **with probability one**) to a random variable $Y : \Omega \to \mathbb{R}$ if

$$\mathbf{P}(\lim_{n\to\infty} Y_n = Y) = 1.$$

That is, $\mathbf{P}(\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}) = 1$

Definition 2.2 (**Convergence in Probability**)**.** We say that a sequence of random variables $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ **converges in probability** to a random variable $Y : \Omega \to \mathbb{R}$ if: for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}(|Y_n - Y| > \varepsilon) = 0.$$

That is, $\forall\ \varepsilon > 0$, $\lim_{n\to\infty} \mathbf{P}(\omega \in \Omega : |Y_n(\omega) - Y(\omega)| > \varepsilon) = 0$.

Definition 2.3 (**Convergence in Distribution**)**.** We say that real-valued random variables $Y_1, Y_2, \ldots$ **converge in distribution** to a real-valued random variable Y if, for any $t \in \mathbb{R}$ such that $s \mapsto \mathbf{P}(Y \le s)$ is continuous at $s = t$,

$$\lim_{n\to\infty} \mathbf{P}(Y_n \le t) = \mathbf{P}(Y \le t).$$

Note that the random variables are allowed to have different domains.

Definition 2.4 (**Convergence in L_p**)**.** Let $0 < p \le \infty$. We say that random variables $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ **converge in L_p** to $Y : \Omega \to \mathbb{R}$ if $\|Y\|_p < \infty$ and

$$\lim_{n\to\infty} \|Y_n - Y\|_p = 0.$$

(Recall that $\|Y\|_p := (\mathbf{E}\,|Y|^p)^{1/p}$ if $0 < p < \infty$ and $\|X\|_\infty := \inf\{c > 0 : \mathbf{P}(|X| \le c) = 1\}$.)

Exercise 2.5. Let $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ be random variables that converge almost surely to a random variable $Y : \Omega \to \mathbb{R}$. Show that $Y_1, Y_2, \ldots$ converges in probability to Y in the following way.

- For any $\varepsilon > 0$ and for any positive integer n, let

$$A_{n,\varepsilon} := \bigcup_{m=n}^{\infty} \{\omega \in \Omega : |Y_m(\omega) - Y(\omega)| > \varepsilon\}.$$

  Show that $A_{n,\varepsilon} \supseteq A_{n+1,\varepsilon} \supseteq A_{n+2,\varepsilon} \supseteq \cdots$.
- Show that $\mathbf{P}(\cap_{n=1}^{\infty} A_{n,\varepsilon}) = 0$.

- Using Continuity of the Probability Law, deduce that $\lim_{n\to\infty} \mathbf{P}(A_{n,\varepsilon}) = 0$.

Now, show that the converse is false. That is, find random variables $Y_1, Y_2, \ldots$ that converge in probability to Y, but where $Y_1, Y_2, \ldots$ do not converge to Y almost surely.

Exercise 2.6. Let $0 < p \leq \infty$. Show that, if $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ converge to $Y : \Omega \to \mathbb{R}$ in L_p, then $Y_1, Y_2, \ldots$ converges to Y in probability.

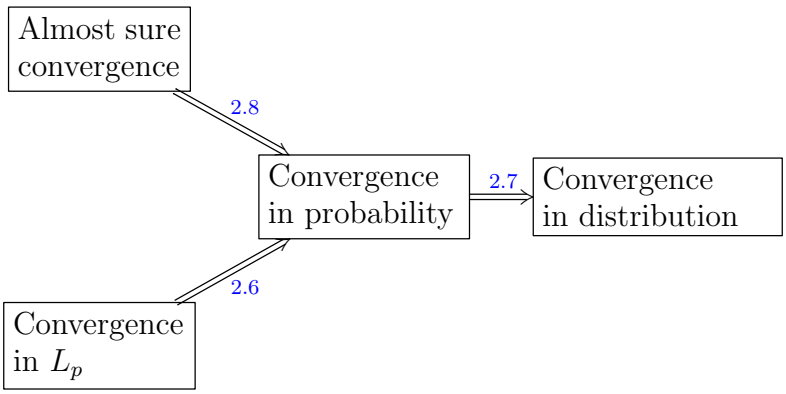
Then, show that the converse is false.

Exercise 2.7. Suppose random variables $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ converge in probability to a random variable $Y : \Omega \to \mathbb{R}$. Prove that $Y_1, Y_2, \ldots$ converge in distribution to Y.

Then, show that the converse is false.

Exercise 2.8. Prove the following statement. Almost sure convergence does not imply convergence in L_2, and convergence in L_2 does not imply almost sure convergence. That is, find random variables that converge in L_2 but not almost surely. Then, find random variables that converge almost surely but not in L_2.

Remark 2.9. The following table summarizes our different notions of convergence of random variables, i.e. the following table summarizes the implications of Exercises 2.6, 2.7 and 2.8.



2.2. **Limit Theorems.** Laws of Large numbers say that if you perform a poll, then the sample mean converges to the mean of the random variable, *regardless of the population size.* Or, in the terminology of elementary statistics, the sample mean becomes more accurate as the sample size increases. We will discuss the sample mean and related concepts more in Section 4.

Theorem 2.10 (**Weak Law of Large Numbers**). *Let $X_1, \ldots, X_n$ be independent identically distributed random variables. Assume that $\mu := \mathbf{E}X_1$ is finite. Then for any $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbf{P}\left(\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| > \varepsilon \right) = 0.$$

Theorem 2.11 (**Strong Law of Large Numbers**). *Let $X_1, \ldots, X_n$ be independent identically distributed random variables. Assume that $\mu := \mathbf{E}X_1$ is finite. Then*

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu \right) = 1.$$

Remark 2.12. A Monte Carlo simulation takes n independent samples from some random distribution and then sums the sample results and divides by n. The Strong Law of

Large Numbers guarantees that this averaging procedure converges to the average value as n becomes large.

The Laws of Large Numbers unfortunately say nothing about the distribution of the sum $X_1 + \cdots + X_n$. Or, in the terminology of elementary statistics, the precision of the sample mean is not addressed by the Laws of Large Numbers. The precision of the sum $X_1 + \cdots + X_n$ is instead dealt with in the Central Limit Theorem. This Theorem was apparently called "Central" since it is so fundamental to probability and statistics, and mathematics more generally.

More formally, let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be i.i.d. random variables with mean zero and variance 1. From the Strong Laws of Large Numbers, $\frac{1}{n}(X_1 + \cdots + X_n)$ converges to 0 almost surely (and in probability). From these results, it is still unclear what value $X_1 + \cdots + X_n$ "typically" takes. For example, if $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = -1) = 1/2$, then $\lim_{n\to\infty} \mathbf{P}(X_1 + \cdots + X_n = 0) = 0$. (What is the exact probability that $\mathbf{P}(X_1 + \cdots + X_n = 0)$?) In order to see what values $X_1 + \cdots + X_n$ "typically" takes, we need to divide by a constant smaller than $\sqrt{n \log n}$

Consider $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$. Dividing by $\sqrt{n}$ is quite natural since $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ has mean zero and variance 1 by Exercise 1.58. So, we expect that the most typical values of $X_1 + \cdots + X_n$ occur in some range $(-a\sqrt{n}, a\sqrt{n})$ for some $a > 0$.

Dividing by anything other than $\sqrt{n}$ will not work correctly. For example, if $g\colon \mathbb{N} \to (0, \infty)$ satisfies $\lim_{n\to\infty} g(n) = \infty$, then it follows from Chebyshev's inequality, Corollary 1.97, that $\frac{1}{g(n)\sqrt{n}}(X_1 + \cdots + X_n)$ converges to 0 in probability. Similarly, $\frac{g(n)}{\sqrt{n}}(X_1 + \cdots + X_n)$ does not converge in any sensible way as $n \to \infty$ (though we will not show this here). In summary, in order to see what values $X_1 + \cdots + X_n$ typically takes, we must divide by $\sqrt{n}$.

Unfortunately, we cannot hope for $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ to converge almost surely or in probability. (We will not show this here.) So, we have to look for a different notion of convergence.

Theorem 2.13 (Central Limit Theorem). *Let $X_1, \ldots, X_n$ be independent identically distributed random variables. Assume that $\mathbf{E}\,|X_1| < \infty$ and $0 < \mathrm{Var}(X_1) < \infty$.*

Let $\mu = \mathbf{E}X_1$ and let $\sigma = \sqrt{\mathrm{Var}(X_1)}$. Then for any $-\infty \le a \le \infty$,

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - \mu n}{\sigma\sqrt{n}} \le a\right) = \int_{-\infty}^{a} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}.$$

Remark 2.14. The random variable $\frac{X_1 + \cdots + X_n - (1/2)n}{\sigma\sqrt{n}}$ has mean zero and variance 1, just like the standard Gaussian.

Exercise 2.15. Estimate the probability that 1000000 coin flips of fair coins will result in more than 501,000 heads, using the Central Limit Theorem. (Some of the following integrals may be relevant: $\int_{-\infty}^{0} e^{-t^2/2} dt/\sqrt{2\pi} = 1/2$, $\int_{-\infty}^{1} e^{-t^2/2} dt/\sqrt{2\pi} \approx .8413$, $\int_{-\infty}^{2} e^{-t^2/2} dt/\sqrt{2\pi} \approx .9772$, $\int_{-\infty}^{3} e^{-t^2/2} dt/\sqrt{2\pi} \approx .9987$.) (Hint: use Bernoulli random variables.)

Casinos do these kinds of calculations to make sure they make money and that they do not go bankrupt. Financial institutions and insurance companies do similar calculations for similar reasons.

Exercise 2.16. Let X, Y be independent, discrete random variables. Using a total probability theorem-type argument, show that

$$\mathbf{P}(X + Y = z) = \sum_{x \in \mathbb{R}} \mathbf{P}(X = x)\mathbf{P}(Y = z - x), \qquad \forall\, z \in \mathbb{R}.$$

Exercise 2.17. Let X, Y be independent, continuous random variables with densities f_X, f_Y, respectively. Let f_{X+Y} be the density of $X + Y$. Show that

$$f_{X+Y}(z) = \int_{\mathbb{R}} f_X(x) f_Y(z - x) dx, \qquad \forall z \in \mathbb{R}.$$

Using this identity, find the density f_{X+Y} when X and Y are both independent, uniformly distributed on $[0, 1]$.

Exercise 2.18 (**Confidence Intervals**)**.** Among 625 members of a bank chosen uniformly at random among all bank members, it was found that 25 had a savings account. Give an interval of the form $[a, b]$ where $0 \leq a, b \leq 625$ are integers, such that with about 95% certainty, the number of any set of 625 bank members with savings accounts chosen uniformly at random lies in the interval $[a, b]$. (Hint: if Y is a standard Gaussian random variable, then $\mathbf{P}(-2 \leq Y \leq 2) \approx .95$.)

Exercise 2.19 (**Hypothesis Testing**)**.** Suppose we run a casino, and we want to test whether or not a particular roulette wheel is biased. Let p be the probability that red results from one spin of the roulette wheel. Using statistical terminology, "$p = 18/38$" is the null hypothesis, and "$p \neq 18/38$" is the alternative hypothesis. (On a standard roulette wheel, 18 of the 38 spaces are red.) For any $i \geq 1$, let $X_i = 1$ if the i^{th} spin is red, and let $X_i = 0$ otherwise.

Let $\mu := \mathbf{E}X_1$ and let $\sigma := \sqrt{\text{var}(X_1)}$. If the null hypothesis is true, and if Y is a standard Gaussian random variable

$$\lim_{n \to \infty} \mathbf{P}\left(\left| \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \right| \geq 2 \right) = \mathbf{P}(|Y| \geq 2) \approx .05.$$

To test the null hypothesis, we spin the wheel n times. In our test, we reject the null hypothesis if $|X_1 + \cdots + X_n - n\mu| > 2\sigma\sqrt{n}$. Rejecting the null hypothesis when it is true is called a type I error. In this test, we set the type I error percentage to be 5%. (The type I error percentage is closely related to the p-value.)

Suppose we spin the wheel $n = 3800$ times and we get red 1868 times. Is the wheel biased? That is, can we reject the null hypothesis with around 95% certainty?

Exercise 2.20 (**Numerical Integration**)**.** In computer graphics in video games, etc., various integrations are performed in order to simulate lighting effects. Here is a way to use random sampling to integrate a function in order to quickly and accurately render lighting effects. Let $\Omega = [0, 1]$, and let $\mathbf{P}$ be the uniform probably law on Ω, so that if $0 \leq a < b \leq 1$, we have $\mathbf{P}([a, b]) = b - a$. Let $X_1, \ldots, X_n$ be independent random variables such that $\mathbf{P}(X_i \in [a, b]) = b - a$ for all $0 \leq a < b \leq 1$, for all $i \in \{1, \ldots, n\}$. Let $f \colon [0, 1] \to \mathbb{R}$ be a continuous function we would like to integrate. Instead of integrating f directly, we instead compute the quantity

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Show that

$$\lim_{n\to\infty} \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) = \int_0^1 f(t)dt.$$

$$\lim_{n\to\infty} \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) = 0.$$

That is, as n becomes large, $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ is a good estimate for $\int_0^1 f(t)dt$.

Exercise 2.21 (Optional; Numerical Integration, Continued)**.** Let $\mathbf{P}$ denote the uniform probability law on $[0,1]$, and let $X\colon [0,1] \to \mathbb{R}$ be a random variable. This exercise discusses how to numerically compute expected values on a computer, as in Exercise 2.20. The procedure below is an example of **Monte Carlo simulation**.

Consider the function $X(t) := t$ for all $t \in [0,1]$. We know that $\mathbf{E}X = 1/2$. To approximate $\mathbf{E}X$ with Matlab, we can use `sum(rand(1,1000))/1000`, which sums 1000 independent, random samples from the uniform probability law on $[0,1]$, and averages them (by dividing by 1000). Enter the term `sum(rand(1,1000))/1000` a few times in the command line of Matlab, to get a few different results.

Consider the function $X(t) := t^2$ for all $t \in [0,1]$. Using Matlab, approximate $\mathbf{E}X$ by averaging 1000 random samples from the uniform probability law on $[0,1]$.

Now, let $\mathbf{P}$ denote the standard Gaussian probability law on $\mathbb{R}$, so that

$$\mathbf{E}X := \int_{-\infty}^{\infty} X(t)e^{-t^2/2}dt/\sqrt{2\pi}$$

for any function $X\colon \mathbb{R} \to \mathbb{R}$. Using the Matlab function `randn`, approximate $\mathbf{E}X$ for $X(t) := t$ and $X(t) := t^2$ by averaging 1000 random samples from the standard Gaussian probability law.

Remark 2.22. When Matlab or other computer programs generate "random numbers" using e.g. `rand` or `randn`, these numbers are not actually random or independent. These numbers are **pseudorandom**. That is, functions such as `rand` output numbers in a deterministic way, but these numbers behave as if they were random. All "random" numbers generated by computers are actually pseudorandom, and this includes slot machines at casinos, video games, etc. So, when using Monte Carlo simulation as we did above, we should be careful about interpreting our results, since it is generally impossible to take random samples from a probability law on a computer.

And, theoretically, if you knew enough about the random number generator that a slot machine is using, you could predict its output.

Exercise 2.23. Suppose you begin at the lower left corner of an 8×8 chess board. Every day, you are allowed to move either up or right to a consecutive board space (unless you are waiting). When you land on a new space, you have to wait a number of days specified by the number sitting on that board space, until you move again. The numbers on the board

spaces appear below.

$$\begin{pmatrix} 1 & 2 & 2 & 1 & 3 & 2 & 6 & 0 \\ 4 & 7 & 3 & 2 & 4 & 8 & 3 & 4 \\ 3 & 4 & 4 & 4 & 5 & 5 & 4 & 2 \\ 4 & 7 & 5 & 3 & 4 & 4 & 5 & 5 \\ 4 & 5 & 4 & 2 & 3 & 3 & 7 & 3 \\ 4 & 6 & 6 & 4 & 3 & 4 & 3 & 2 \\ 5 & 4 & 6 & 3 & 4 & 3 & 4 & 1 \\ 0 & 3 & 6 & 2 & 7 & 2 & 7 & 5 \end{pmatrix}.$$

Your goal is to reach the top right corner of the chess board in the shortest amount of time. Find the path that takes the shortest amount of time, and also find the shortest amount of time that it takes to reach the top right corner. (Hint: Use recursion. That is, solve a more general problem. For *any* square on the board, find the least number of days it takes to reach that square starting from the bottom left corner, using only up and right moves. If you are still stuck, read a bit about dynamic programming.)

Exercise 2.24 (**Renewal Theory**)**.** Let $t_1, t_2, \ldots$ be positive, independent identically distributed random variables. Let $\mu \in \mathbb{R}$. Assume $\mathbf{E}t_1 = \mu$. For any positive integer j, we interpret t_j as the lifetime of the j^{th} lightbulb (before burning out, at which point it is replaced by the $(j+1)^{st}$ lightbulb). For any $n \geq 1$, let $T_n := t_1 + \cdots + t_n$ be the total lifetime of the first n lightbulbs. For any positive integer t, let $N_t := \min\{n \geq 1 \colon T_n \geq t\}$ be the number of lightbulbs that have been used up until time t. Show that N_t/t converges almost surely to $1/\mu$ as $t \to \infty$. (Hint: if c, t are positive integers, then $\{N_t \leq ct\} = \{T_{ct} \geq t\}$. Apply the Strong Law to T_{ct}.)

Exercise 2.25 (**Playing Monopoly Forever**)**.** Let $t_1, t_2, \ldots$ be independent random variables, all of which are uniform on $\{1, 2, 3, 4, 5, 6\}$. For any positive integer j, we think of t_j as the result of rolling a single fair six-sided die. For any $n \geq 1$, let $T_n = t_1 + \cdots + t_n$ be the total number of spaces that have been moved after the n^{th} roll. (We think of each roll as the amount of moves forward of a game piece on a very large Monopoly game board.) For any positive integer t, let $N_t := \min\{n \geq 1 \colon T_n \geq t\}$ be the number of rolls needed to get t spaces away from the start. Using Exercise 2.24, show that N_t/t converges almost surely to $2/7$ as $t \to \infty$.

Exercise 2.26 (**Random Numbers are Normal**)**.** Let X be a uniformly distributed random variable on $(0, 1)$. Let X_1 be the first digit in the decimal expansion of X. Let X_2 be the second digit in the decimal expansion of X. And so on.

- Show that the random variables $X_1, X_2, \ldots$ are uniform on $\{0, 1, 2, \ldots, 9\}$ and independent.
- Fix $m \in \{0, 1, 2, \ldots, 9\}$. Using the Strong Law of Large Numbers, show that with probability one, the fraction of appearances of the number m in the first n digits of X converges to $1/10$ as $n \to \infty$.

(Optional): Show that for any ordered finite set of digits of length k, the fraction of appearances of this set of digits in the first n digits of X converges to 10^{-k} as $n \to \infty$. (You already proved the case $k = 1$ above.) That is, a randomly chosen number in $(0, 1)$ is normal. On the other hand, if we just pick some number such that $\sqrt{2} - 1$, then it may not be easy to say whether or not that number is normal.

(As an optional exercise, try to explicitly write down a normal number. This may not be so easy to do, even though a random number in $(0,1)$ satisfies this property!)

2.3. **Additional Comments.** A version of the Law of Large Numbers was stated as early as the 1500s. In the 1700s and 1800s, various laws of large numbers were proved with weaker and weaker hypotheses. For example, the L_2 Weak Law was known to Chebyshev in 1867. The Strong Law of Large Numbers might have first been proven in 1930 by Kolmogorov.

If the random variables have infinite mean, then the Strong Law cannot hold.

Exercise 2.27. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be i.i.d. with $\mathbf{E}|X_1| = \infty$. Then $\mathbf{P}(|X_n| > n \text{ for infinitely many } n \geq 1) = 1$. And $\mathbf{P}(\lim_{n\to\infty} \frac{X_1+\cdots+X_n}{n} \in (-\infty, \infty)) = 0$. (Hint: show $\sum_{n=1}^{\infty} \mathbf{P}(|X_n| > n) = \infty$, then apply the second Borel-Cantelli Lemma. Write $\frac{S_n}{n} - \frac{S_{n+1}}{n+1} = \frac{S_n}{n(n+1)} - \frac{X_{n+1}}{n+1}$, and consider what happens to both sides on the set where $\lim_{n\to\infty} \frac{S_n}{n} \in \mathbb{R}$.)

Exercise 2.28 (**Second Borel-Cantelli Lemma**)**.** Let $A_1, A_2, \ldots$ be independent events with $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$. Then $\mathbf{P}(A_n \text{ occurs for infinitely many } n \geq 1) = 1$. (Hint: using $1 - x \leq e^{-x}$ for any $x \in \mathbb{R}$, show $\mathbf{P}(\cap_{n=s}^{t} A_n^c) \leq \exp(-\sum_{n=s}^{t} \mathbf{P}(A_n))$, let $t \to \infty$ to conclude $\mathbf{P}(\cup_{n=s}^{\infty} A_n) = 1$ for all $s \geq 1$, then let $s \to \infty$.)

The Central Limit Theorem was described by de Moivre in 1733 and again by Laplace in 1785 and 1812, where the Fourier Transform was used. In 1901, Lyapunov proved the Central Limit Theorem under an assumption similar to $\mathbf{E}|X_1|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$. The Central Limit Theorem under the assumption of a finite (truncated) second moment was proven by Lindeberg in 1920. This result was extended by Feller in 1935, also with contributions by Lévy in the same year.

Theorem 2.29 (**Lindeberg Central Limit Theorem for Triangular Arrays**)**.** *Let $j_1, j_2, \ldots$ be a sequence of natural numbers with $\lim_{n\to\infty} j_n = \infty$. For any $n \geq 1$, let $X_{n,1}, \ldots, X_{n,n} : \Omega_n \to \mathbb{R}$ be independent with mean zero and finite variance. (Note e.g. that $X_{3,1}$ and $X_{2,2}$ might not be independent, and the sample space is allowed to change as n changes.) Define*

$$\sigma_n^2 := \sum_{k=1}^{j_n} \mathrm{Var}(X_{n,k}), \qquad \forall\, n \geq 1.$$

Assume that $\sigma_n > 0$ for all $n \geq 1$. If, for any $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} \sum_{k=1}^{j_n} \mathbf{E}(|X_{n,k}|^2 1_{|X_{n,k}| > \varepsilon\sigma_n}) = 0, \qquad (*)$$

then the random variables $\frac{X_{n,1}+\cdots+X_{n,n}}{\sigma_n}$ converge in distribution to a standard Gaussian random variable.

The Lindeberg condition $(*)$ implies the Feller condition

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} \max_{1\leq k\leq n} \mathbf{E}|X_{n,k}|^2 = 0.$$

It was shown by Feller that if the above assumptions hold (without $(*)$) and if the Feller condition holds, then the Lindeberg condition $(*)$ is necessary and sufficient for $\frac{X_{n,1}+\cdots+X_{n,n}}{\sigma_n}$

to converge in distribution to a standard Gaussian random variable. The combined result is sometimes known as the Lindeberg-Feller theorem.

Berry and Esseén separately gave an error bound for the Central Limit Theorem in the early 1940s.

Theorem 2.30 (**Berry-Esseén**). *There exists $c > 0$ such that the following holds. Let $X_1, X_2, \dots$ be i.i.d. real-valued random variables with mean zero, variance 1 and $\mathbf{E}|X_1|^3 < \infty$. Let Z be a standard Gaussian random variable. Then for any $n \geq 1$,*

$$\sup_{t \in \mathbb{R}} \left| \mathbf{P}(X_1 + \cdots + X_n/\sqrt{n} < t) - \mathbf{P}(Z < t) \right| \leq c \cdot \frac{\mathbf{E}|X_1|^3}{\sqrt{n}}.$$

With the assumption of more bounded moments, an asymptotic expansion can be written, with explicit dependence on t, for the difference $|\mathbf{P}(X_1 + \cdots + X_n/\sqrt{n} < t) - \mathbf{P}(Z < t)|$. This expansion is called the Edgeworth Expansion; see Feller, Vol. 2, XVI.4.(4.1).

One may ask for general conditions under which the average of any i.i.d. random variables have a limiting distribution, with moment assumptions different than the Central Limit Theorem. Necessary and sufficient conditions are described in the following Theorem.

Theorem 2.31. *Let $X_1, X_2, \dots$ be i.i.d. real-valued random variables. Assume there exists a function $h\colon [0, \infty) \to (0, \infty)$ such that, for any $x > 0$, $\lim_{x \to \infty} L(tx)/L(x) = 1$. Assume also there exists $\theta \in [0, 1]$ and $\alpha \in (0, 2)$such that*

- $\lim_{x \to \infty} \mathbf{P}(X_1 > x)/\mathbf{P}(|X_1| > x) = \theta$,
- $\mathbf{P}(|X_1| > x) = x^{-\alpha} L(x)$, $\forall\ x > 0$.

For any $n \geq 1$, define

$$a_n := \inf\{x > 0\colon P(|X_1| > x) \leq 1/n\}, \qquad b_n := \mathbf{E}(X_1 1_{|X_1| \leq a_n}).$$

Then $\frac{X_1 + \cdots + X_n - a_n}{b_n}$ converges in distribution to a random variable Y as $n \to \infty$

Exercise 2.32. Show that there exists a nonzero random variable X such that, if $X_1, X_2, \dots$ are i.i.d. copies of X, then $\frac{X_1 + \cdots + X_n}{n}$ is equal in distribution to X, for any $n \geq 1$. (Optional: can you write out an explicit formula for the density of X?) (Hint: take the Fourier transform.)

Show that there exists a nonzero random variable X such that, if $X_1, X_2, \dots$ are i.i.d. copies of X, then $\frac{X_1 + \cdots + X_n}{n^2}$ is equal in distribution to X, for any $n \geq 1$.

By projection the random variables onto one-dimensional lines, the following Central Limit Theorem in $\mathbb{R}^d$ can be proven from the corresponding result in $\mathbb{R}$.

Theorem 2.33 (**Central Limit Theorem in $\mathbb{R}^d$**). *Let $X^{(1)}, X^{(2)}, \dots$ be i.i.d. $\mathbb{R}^d$-valued random variables. Let $\mu \in \mathbb{R}^d$. (We write a random variable in its components as $X^{(n)} = (X_1^{(n)}, \dots, X_d^{(n)}) \in \mathbb{R}^d$.) Assume $\mathbf{E}X^{(n)} = \mu$ for all $n \geq 1$, and for any $1 \leq i, j \leq d$, all of the covariances*

$$a_{ij} := \mathbf{E}((X_i^{(1)} - \mathbf{E}X_i^{(1)})(X_j^{(1)} - \mathbf{E}X_j^{(1)})).$$

are finite. Then as $n \to \infty$, $\frac{X^{(1)} + \cdots + X^{(n)} - n\mu}{\sqrt{n}}$ converges weakly to a Gaussian random vector $Z = (Z_1, \dots, Z_d) \in \mathbb{R}^d$ with covariance matrix $(a_{ij})_{1 \leq i,j \leq d}$.

Remark 2.34. By definition, a random vector $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ is **Gaussian** if, for any $v_1, \ldots, v_d \in \mathbb{R}$, the random variable $\sum_{i=1}^d v_i Z_i$ is a Gaussian random variable. Equivalently, for any $v \in \mathbb{R}^d$, the random variable $\langle v, Z \rangle$ is a Gaussian random variable. The covariance matrix $(a_{ij})_{1 \leq i,j \leq d}$ of Z is defined by

$$a_{ij} := \mathbf{E}((Z_i - \mathbf{E}Z_i)(Z_j - \mathbf{E}Z_j)).$$

Exercise 2.35. Let $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ be a Gaussian random vector.

- Show that the covariance matrix $(a_{ij})_{1 \leq i,j \leq d}$ of Z is symmetric, positive semidefinite. That is, for any $v \in \mathbb{R}^d$, we have

$$v^T a v = \sum_{i,j=1}^d v_i v_j a_{ij} \geq 0.$$

- Given any symmetric positive semidefinite matrix $(b_{ij})_{1 \leq i,j \leq d}$, show that there exists a Gaussian random vector Z such that the covariance matrix of Z is $(b_{ij})_{1 \leq i,j \leq d}$. (Hint: write the matrix b in its Cholesky decomposition $b = rr^*$, where r is a $d \times d$ real matrix. Let $e^{(1)}, \ldots, e^{(d)}$ be the rows of r. Let $X_1, \ldots, X_d$ be independent standard Gaussian random variables. Let $X := (X_1, \ldots, X_d)$. Define $Z_i := \langle X, e^{(i)} \rangle$ for any $1 \leq i \leq d$.)

Proposition 2.36.

- *(Slutsky's Theorem) Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables that converge in distribution to $X : \Omega \to \mathbb{R}$. Let $c \in \mathbb{R}$. Let $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ be random variables that converge in probability to c. Then $X_1 + Y_1, X_2 + Y_2, \ldots$ converges in distribution to $X + c$. Also, $X_1 Y_1, X_2 Y_2, \ldots$ converges in distribution to cX.*
- *Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables that converge in distribution to $X : \Omega \to \mathbb{R}$. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous. Then $f(X_1), f(X_2), \ldots$ converges in distribution to $f(X)$.*

Exercise 2.37. Suppose I tell you that the following list of 20 numbers is a random sample from a Gaussian random variable, but I don't tell the mean or standard deviation.

5.1715, 3.2925, 5.2172, 6.1302, 4.9889, 5.5347, 5.2269, 4.1966, 4.7939, 3.7127

5.3884, 3.3529, 3.4311, 3.6905, 1.5557, 5.9384, 4.8252, 3.7451, 5.8703, 2.7885

To the best of your ability, determine what the mean and standard deviation are of this random variable. (This question is a bit open-ended, so there could be more than one correct way of justifying your answer.)

Exercise 2.38. Suppose I tell you that the following list of 20 numbers is a random sample from a Gaussian random variable, but I don't tell you the mean or standard deviation. Also, around one or two of the numbers was corrupted by noise, computational error, tabulation error, etc., so that it is totally unrelated to the actual Gaussian random variable.

−1.2045, −1.4829, −0.3616, −0.3743, −2.7298, −1.0601, −1.3298, 0.2554, 6.1865, 1.2185

−2.7273, −0.8453, −3.4282, −3.2270, −1.0137, 2.0653, −5.5393, −0.2572, −1.4512, 1.2347

To the best of your ability, determine what the mean and standard deviation are of this random variable. Supposing you had instead a billion numbers, and 5 or 10 percent of them were corrupted samples, can you come up with some automatic way of throwing out the corrupted samples? (Once again, there could be more than one right answer here; the question is intentionally open-ended.)

3. Exponential Families

A basic problem in statistics is to fit data to an unknown probability distribution. As in Exercise 2.37, we might have a list of numbers, and we known these numbers follow some Gaussian distribution, but we might not know the mean and variance of this Gaussian. We then want to infer the mean and variance from the data. In this example, there are two unknown parameters. In order to generalize this problem, we introduce exponential families. Exponential families provide a general class of distributions with a given number of unknown parameters. Many of the examples introduced in Section 1.2 can be understood as exponential families.

Definition 3.1 (**Exponential Families**)**.** Let n, k be a positive integers and let μ be a measure on $\mathbb{R}^n$. Let $t_1, \dots, t_k \colon \mathbb{R}^n \to \mathbb{R}$. Let $h \colon \mathbb{R}^n \to [0, \infty)$ with $\mu(\{x \in \mathbb{R}^n \colon h(x) > 0\}) > 0$. For any $w = (w_1, \dots, w_k) \in \mathbb{R}^k$, define

$$a(w) := \log \int_{\mathbb{R}^n} h(x) \exp\Big(\sum_{i=1}^{k} w_i t_i(x) \Big) d\mu(x).$$

The set $\{w \in \mathbb{R}^k \colon a(w) < \infty\}$ is called the **natural parameter space**. On this set, the function

$$f_w(x) := h(x) \exp\Big(\sum_{i=1}^{k} w_i t_i(x) - a(w) \Big), \qquad \forall\, x \in \mathbb{R}^n$$

satisfies $\int_{\mathbb{R}^n} f_w(x) d\mu(x) = 1$. So, the set of functions (which can be interpreted as probability density functions, or as probability mass functions according to μ)

$$\{f_w \colon a(w) < \infty\}$$

is called a **k-parameter exponential family in canonical form**.

More generally, let $\Theta \subseteq \mathbb{R}^k$ and let $w \colon \Theta \to \mathbb{R}^k$. We define a **$k$-parameter exponential family** to be a set of functions $\{f_\theta \colon \theta \in \Theta,\, a(w(\theta)) < \infty\}$, where

$$f_\theta(x) := h(x) \exp\Big(\sum_{i=1}^{k} w_i(\theta) t_i(x) - a(w(\theta)) \Big), \qquad \forall\, x \in \mathbb{R}^n.$$

An exponential family is called **curved** if the dimension of Θ is less than k.

Remark 3.2. If $w \colon \Theta \to \mathbb{R}^k$ has an inverse function, then the corresponding k-parameter exponential family can be written in canonical form.

When we deal with probability density functions, we will simplify to $d\mu(x) = dx$ and $n = 1$, so that

$$a(w) := \log \int_{\mathbb{R}} h(x) \exp\Big(\sum_{i=1}^{k} w_i t_i(x) \Big) dx.$$

and we can then interpret

$$f_\theta(x) := h(x) \exp\Big(\sum_{i=1}^{k} w_i(\theta) t_i(x) - a(w(\theta)) \Big), \qquad \forall\, x \in \mathbb{R}$$

as probability density functions on the real line, since $\int_{\mathbb{R}} f_\theta(x) dx = 1$ for every θ such that $a(w(\theta)) < \infty$, and $f_{w(\theta)}(x) \geq 0$ for all $x \in \mathbb{R}$.

To specialize to probability mass functions on e.g. the integers, we let μ be counting measure (so that $\mu(\{m\}) = 1$ for any integer m, and $\mu(\{x\}) = 0$ for any $x \in \mathbb{R}$ that is not an integer), so that

$$a(w) := \log \sum_{m=-\infty}^{\infty} h(m) \exp\Big(\sum_{i=1}^{k} w_i(\theta) t_i(m) \Big).$$

and we can then interpret

$$f_\theta(m) := h(m) \exp\Big(\sum_{i=1}^{k} w_i(\theta) t_i(m) - a(w(\theta)) \Big), \qquad \forall\, m \in \mathbb{Z}$$

as a probability mass function, since $\sum_{m \in \mathbb{Z}} f_{w(\theta)}(m) = 1$ and $f_{w(\theta)}(m) \geq 0$ for all $m \in \mathbb{Z}$.

Below we will use f_θ interchangeably for a single variable density/mass function and for a joint density/mass function.

Example 3.3. Let us see how to phrase Exercise 2.37 using a two parameter exponential family. We write a Gaussian density of mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$ as

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} \exp\Big(\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \Big(\frac{\mu^2}{2\sigma^2} + \log \sigma \Big) \Big), \qquad \forall\, x \in \mathbb{R}.$$

Then, we interpret θ as $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2) \in \mathbb{R}^2$, and define

$$t_1(x) := x, \qquad t_2(x) := x^2,$$

$$w_1(\theta) := \frac{\theta_1}{\theta_2} = \frac{\mu}{\sigma^2}, \qquad w_2(\theta) := -\frac{1}{2\theta_2} = -\frac{1}{2\sigma^2},$$

$$a(w(\theta)) := \frac{\theta_1^2}{2\theta_2} + \frac{1}{2} \log \theta_2 = \frac{\mu^2}{2\sigma^2} + \log \sigma,$$

and $h(x) := \frac{1}{\sqrt{2\pi}}$ for all $x \in \mathbb{R}$. Let $\Theta := \{\theta \in \mathbb{R}^2 \colon \theta_2 > 0\}$, and for any $\theta \in \Theta$, define

$$f_\theta(x) := h(x) \exp\Big(\sum_{i=1}^{2} w_i(\theta) t_i(x) - a(w(\theta)) \Big), \qquad \forall\, x \in \mathbb{R}.$$

Then $\{f_\theta \colon \theta \in \Theta\}$ is a two parameter exponential family.

If we instead want to write this exponential family in canonical form, we replace the θ terms with w_1, w_2 terms as follows

$$a(w) = \frac{\mu^2}{2\sigma^2} + \log \sigma = \Big(\frac{\mu}{\sigma^2} \Big)^2 \Big[(-4) \frac{(-1)}{2\sigma^2} \Big]^{-1} - \frac{1}{2} \log \Big((-2) \frac{(-1)}{2\sigma^2} \Big) = -\frac{w_1^2}{4w_2} - \frac{1}{2} \log(-2w_2).$$

We then restrict to the set $\{(w_1, w_2) \in \mathbb{R}^2 \colon w_2 < 0\}$ and define

$$f_w(x) := h(x) \exp\Big(\sum_{i=1}^{2} w_i t_i(x) - a(w)) \Big), \qquad \forall\, x \in \mathbb{R}.$$

Remark 3.4 (**Location Family**)**.** Let X be a random variable with density $f \colon \mathbb{R} \to \mathbb{R}$. Let $\mu \in \mathbb{R}$. Then the densities $\{f(x+\mu)\}_{\mu \in \mathbb{R}}$ are called the **location family** of X. This family may or may not be an exponential family.

Exercise 3.5. Let X be uniformly distributed on $[0, 1]$. Show that the location family of X is not an exponential family in the following sense. The corresponding densities $\{f(x+\mu)\}_{\mu \in \mathbb{R}}$ cannot be written in the form

$$h(x) \exp(w(\mu) t(x) - a(w(\mu)))$$

where $h \colon \mathbb{R} \to \mathbb{R}$, $w \colon \mathbb{R} \to \mathbb{R}$, $t \colon \mathbb{R} \to \mathbb{R}$, $x \in \mathbb{R}$ and $a(w(\mu))$ is a real number chosen so that the integral of the density is one. (Hint: Argue by contradiction. Assume that the location family is a one-parameter exponential family. Compare where the different densities are zero or nonzero as the parameter changes.)

Remark 3.6 (**Scale Family**)**.** Let X be a random variable with density $f \colon \mathbb{R} \to \mathbb{R}$. Let $\sigma > 0$. Then the densities $\{\sigma^{-1} f(x/\sigma)\}_{\sigma > 0}$ are called the **scale family** of X. This family may or may not be an exponential family. Note that these are probability densities since $\int_{-\infty}^{\infty} \sigma^{-1} f(x/\sigma) dx = \int_{-\infty}^{\infty} f(x) dx = 1$.

Remark 3.7 (**Location and Scale Family**)**.** Let X be a random variable with density $f \colon \mathbb{R} \to \mathbb{R}$. Let $\mu \in \mathbb{R}, \sigma > 0$. Then the densities $\{\sigma^{-1} f((x+\mu)/\sigma)\}_{\sigma > 0}$ are called the **location and scale family** of X. This family may or may not be an exponential family.

3.1. **Differential Identities.** Recall from Exercise 1.41 that a standard Gaussian random variable X satisfies

$$\mathbf{E} e^{tX} = e^{t^2/2}, \qquad \forall\, t \in \mathbb{R},$$

and using this information we can recover the m^{th} moment of X by the formula

$$\frac{d^m}{dt^m}|_{t=0} \mathbf{E} e^{tX} = \mathbf{E} X^m.$$

Similarly, we can differentiate the parameters of exponential families and find out information about moments of the exponential family. We describe such a procedure below.

As in Definition 3.1, let

$$a(w) := \log \int_{\mathbb{R}^n} h(x) \exp\Big(\sum_{i=1}^{k} w_i t_i(x) \Big) d\mu(x).$$

Define now

$$W := \{w \in \mathbb{R}^k \colon a(w) < \infty\}.$$

Lemma 3.8. *The function $a(w)$ is continuous and has continuous partial derivatives of all orders on the interior of W. Moreover, we can compute these derivatives by differentiating under the integral sign.*

Proof. We prove only the case of a first order partial derivative. Consider the case of the partial derivative with respect to w_1 at w in the interior of W. Let $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^k$. Since the exponential function is analytic, it suffices to show that the partial derivative of $e^{a(w)}$ exists in the direction e_1. We form the difference quotient for $e^{a(w)}$ as follows.

$$\begin{aligned}
&\frac{\exp\Big(a(w+\varepsilon e_1)\Big) - \exp(a(w))}{\varepsilon} \\
&\quad = \frac{1}{\varepsilon}\int_{\mathbb{R}^n} h(x)\Big[\exp\Big(\varepsilon t_1(x) + \sum_{i=1}^k w_i t_i(x)\Big) - \exp\Big(\sum_{i=1}^k w_i t_i(x)\Big)\Big] d\mu(x) \\
&\quad = \int_{\mathbb{R}^n} h(x)\frac{\exp(\varepsilon t_1(x)) - 1}{\varepsilon}\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big) d\mu(x).
\end{aligned}$$

By the Mean Value Theorem, for any $0 < \alpha < 1$ and for any $\beta \in \mathbb{R}$

$$\left|e^{\alpha\beta} - 1\right| \le |\alpha\beta| \max(1, e^{\alpha\beta}) \le |\alpha\beta|\, e^{|\beta|} \le |\alpha|\, e^{2|\beta|} \le |\alpha|\,(e^{2\beta} + e^{-2\beta}), \qquad (*)$$

So, using $\delta > 0$, $\alpha := \varepsilon/\delta$ and $\beta := \delta t_1(x)$

$$\begin{aligned}
&\left|h(x)\frac{\exp(\varepsilon t_1(x)) - 1}{\varepsilon}\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big)\right| \\
&\qquad \le h(x)\left|\frac{\exp(\varepsilon t_1(x)) - 1}{\varepsilon}\right| \exp\Big(\sum_{i=1}^k w_i t_i(x)\Big) d\mu(x) \\
&\qquad \overset{(*)}{\le} \frac{1}{\delta} h(x)\Big(e^{2\delta t_1(x)} + e^{-2\delta t_1(x)}\Big)\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big) d\mu(x)
\end{aligned}$$

So, if

$$X_\varepsilon := h(x)\frac{\exp(\varepsilon t_1(x)) - 1}{\varepsilon}\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big),$$

$$Y := \frac{1}{\delta} h(x)\Big(e^{2\delta t_1(x)} + e^{-2\delta t_1(x)}\Big)\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big),$$

then $|X_\varepsilon| \le Y$ for any $0 < \varepsilon < \delta < 1$. We then conclude by the Dominated Convergence Theorem 3.10 that

$$\begin{aligned}
\frac{\partial}{\partial w_1} e^{a(w)} &= \lim_{\varepsilon \to 0} \int_{\mathbb{R}^n} \left| h(x)\frac{\exp(\varepsilon t_1(x)) - 1}{\varepsilon}\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big)\right| d\mu(x) \\
&= \int_{\mathbb{R}^n} t_1(x) h(x) \exp\Big(\sum_{i=1}^k w_i t_i(x)\Big) d\mu(x).
\end{aligned}$$

Here we also use that $\int_{\mathbb{R}^n} Y(x) d\mu(x) = e^{a(w+2\delta e_1)} + e^{a(w-2\delta e_1)} < \infty$ for sufficiently small δ (depending only on w), since w is in the interior of W.

Using the right part of inequality $(*)$, we can similarly show that

$$\int_{\mathbb{R}^n} \prod_{j=1}^{k} |t_j(x)|^{m_j} \, h(x) \exp\Big(\sum_{i=1}^{k} w_i t_i(x) \Big) d\mu(x) < \infty,$$

for any positive integers $m_1, \ldots, m_k$, so that an inductive argument completes the above proof for any iterated partial derivative. □

Remark 3.9. Using Definition 3.1 we can rewrite the penultimate formula as

$$e^{-a(w)} \frac{\partial}{\partial w_1} e^{a(w)} = \int_{\mathbb{R}^n} t_1(x) h(x) \exp\Big(\sum_{i=1}^{k} w_i t_i(x) - a(w) \Big) d\mu(x) = \int_{\mathbb{R}^n} t_1(x) f_w(x) d\mu(x).$$

Theorem 3.10 (Dominated Convergence Theorem). *Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables that converge almost surely. Assume that Y is a nonnegative random variable with $\mathbf{E}Y < \infty$ and $|X_n| \leq Y$ almost surely, $\forall\, n \geq 1$. Then*

$$\mathbf{E} \lim_{n\to\infty} X_n = \lim_{n\to\infty} \mathbf{E} X_n.$$

Corollary 3.11. *Let $\varepsilon > 0$. Let $X \colon \Omega \to \mathbb{R}$ be a random variable such that $\mathbf{E}e^{wX} < \infty$ for all $w \in (-\varepsilon, \varepsilon)$. Then, for any integer $n \geq 1$, $\mathbf{E}X^n$ exists and*

$$\frac{d^n}{dw^n}\Big|_{w=0} \mathbf{E} e^{wX} = \mathbf{E} X^n.$$

Proof. Apply Lemma 3.8 when $\mu = \mathbf{P}$, $h = 1$, $k = 1$, $t(x) = x$. □

Remark 3.12. If X is a Cauchy distributed random variable, then $\mathbf{E}e^{wX} = \infty$ for all $w \neq 0$. So, in this case, the hypothesis of the above Corollary does not apply. Indeed, $\mathbf{E}\,|X| = \infty$ in this case.

Example 3.13. Recall that when $d\mu(x) = dx$ and $n = 1$, we have

$$a(w(\theta)) := \log \int_{\mathbb{R}} h(x) \exp\Big(\sum_{i=1}^{k} w_i(\theta) t_i(x) \Big) dx.$$

so that

$$f_\theta(x) := h(x) \exp\Big(\sum_{i=1}^{k} w_i(\theta) t_i(x) - a(w(\theta)) \Big), \qquad \forall\, x \in \mathbb{R}$$

is probability density functions on the real line. If $\mathbf{E}_\theta$ denotes the expected value with respect to this density function, then we have by the chain rule and Remark 3.9

$$e^{-a(w(\theta))} \frac{\partial}{\partial \theta_1} e^{a(w(\theta))} = e^{-a(w(\theta))} \sum_{i=1}^{k} \frac{\partial e^{a(w)}}{\partial w_i} \frac{\partial w_i}{\partial \theta_1} = \sum_{i=1}^{k} \frac{\partial w_i}{\partial \theta_1} \mathbf{E}_\theta t_i = \mathbf{E}_\theta \Big(\sum_{i=1}^{k} \frac{\partial w_i}{\partial \theta_1} t_i \Big).$$

Returning to Example 3.3, where we wrote the Gaussian density of mean μ and standard deviation $\sigma > 0$ as a two-parameter exponential family, we had $k = 2$, $n = 1$, we wrote θ as $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2) \in \mathbb{R}^2$, and define

$$t_1(x) := x, \qquad t_2(x) := x^2,$$
$$w_1(\theta) := \frac{\theta_1}{\theta_2} = \frac{\mu}{\sigma^2}, \qquad w_2(\theta) := -\frac{1}{2\theta_2} = -\frac{1}{2\sigma^2},$$

$$a(w(\theta)) := \frac{\theta_1^2}{2\theta_2} + \frac{1}{2}\log\theta_2 = \frac{\mu^2}{2\sigma^2} + \log\sigma,$$

So that

$$e^{-a(w(\theta))}\frac{\partial}{\partial\theta_1}e^{a(w(\theta))} = \frac{\theta_1}{\theta_2} = \frac{\mu}{\sigma^2} = \mathbf{E}_\theta\Big(\frac{1}{\theta_2}x\Big) = \frac{1}{\sigma^2}\mathbf{E}_\theta(x).$$

That is, $\mathbf{E}_\theta(x) = \mu$, as we anticipated.

Exercise 3.14. Using a two parameter exponential family for a Gaussian random variable (with mean μ and variance σ^2), compute both sides of the following identity in terms of μ and σ:

$$e^{-a(w)}\frac{\partial^2}{\partial w_i\partial w_j}e^{a(w)} = \int_{\mathbb{R}} t_i(x)t_j(x)h(x)\exp\Big(\sum_{i=1}^{2} w_i t_i(x) - a(w)\Big)d\mu(x), \qquad \forall\, 1 \le i,j \le 2.$$

Recall that in this case,

$$t_1(x) := x, \qquad t_2(x) := x^2, \qquad w_1 := \frac{\mu}{\sigma^2}, \qquad w_2 := -\frac{1}{2\sigma^2},$$
$$a(w) := -\frac{w_1^2}{4w_2} - \frac{1}{2}\log(-2w_2).$$

Example 3.15. We write the binomial distribution with parameters n, p as a one parameter exponential family (where n is fixed), and then take derivatives to find its moments. Recall from Definition 1.19 that the binomial random variable X with parameters n, p satisfies: if x is an integer with $0 \le x \le n$, then

$$\mathbf{P}(X = x) = \binom{n}{x}p^x(1-p)^{n-x}.$$

For any other x, we have $\mathbf{P}(X = x) = 0$. We write

$$\binom{n}{x}p^x(1-p)^{n-x} = \binom{n}{x}(1-p)^n\Big(\frac{p}{1-p}\Big)^x = \binom{n}{x}\exp\Big(x\log\Big(\frac{p}{1-p}\Big) - (-1)n\log(1-p)\Big)$$

So, define $h\colon \mathbb{R} \to \mathbb{R}$ so that, if x is an integer with $0 \le x \le n$, then $h(x) = \binom{n}{x}$. (It does not matter what other values h takes in this example.) Let $\theta := p$, $\Theta := (0,1)$ and define

$$t(x) := x, \qquad w(\theta) := \log\Big(\frac{\theta}{1-\theta}\Big), \qquad a(w(\theta)) := -n\log(1-\theta).$$

Since X only takes values integers values with positive probability, we have

$$f_\theta(x) := h(x)\exp\Big(w(\theta)t(x) - a(w(\theta))\Big), \qquad \forall\, x \in \mathbb{R}$$

As in the previous example, we have

$$e^{-a(w(\theta))}\frac{d}{d\theta}e^{a(w(\theta))} = \frac{d}{d\theta}a(w(\theta)) = \mathbf{E}_\theta\Big(\frac{d}{d\theta}w(\theta)t\Big).$$

That is,

$$\Big(\frac{1}{\theta} + \frac{1}{1-\theta}\Big)\mathbf{E}_\theta(x) = \frac{n}{1-\theta}.$$

That is, the expected value of the binomial with parameters n, p (with $p = \theta$) is

$$\frac{n}{1-\theta}\frac{1}{1/(\theta(1-\theta))} = n\theta.$$

Exercise 3.16. Let $X\colon \Omega \to \mathbb{R}^n$ be a random variable with the **standard Gaussian distribution**:

$$\mathbf{P}(X \in A) := \int_A e^{-(x_1^2+\cdots+x_n^2)/2} dx (2\pi)^{-n/2}, \qquad \forall\, A \subseteq \mathbb{R}^n \text{ measurable.}$$

Let $v_1, \ldots, v_m$ be vectors in $\mathbb{R}^n$. Let $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the standard inner product on $\mathbb{R}^n$, so that $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$ for any $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{R}^n$.

First, let $v \in \mathbb{R}^n$ and show that $\langle X, v \rangle$ is a mean zero Gaussian with variance $\langle v, v \rangle$.

Then, show that the random variables $\langle X, v_1 \rangle, \ldots, \langle X, v_m \rangle$ are independent if and only if the vectors $v_1, \ldots, v_m$ are pairwise orthogonal.

(Hint: use the rotation invariance of the Gaussian.)

3.2. **Additional Comments.** Exponential families were apparently introduced by Darmois, Koopman and Pitman in the 1930s.

4. Random Samples

When conducting a poll of a sample population, one often assumes that there exists a random variable $X\colon \Omega \to \mathbb{R}$ that describes a single observation from the population. Repeated observations of the population are then performed independently of each other. This concept is formalized as a random sample.

Definition 4.1 (**Random Sample**)**.** Let n be a positive integer. A **random sample** of size n is a sequence $X_1, \ldots, X_n$ of independent, identically distributed (i.i.d.) random variables.

As in Exercise 2.37, a basic problem is to find e.g. the mean or standard deviation of the unknown distribution of X. That is, if we have a random sample of size n then $\frac{1}{n}(X_1 + \cdots + X_n)$ seems to be a reasonable guess for the mean of the unknown distribution if n is large. More generally, any function of the random sample is called a statistic.

Definition 4.2 (**Statistic**)**.** Let n, k be positive integers. Let $X_1, \ldots, X_n$ be a random sample of size n. Let $f\colon \mathbb{R}^n \to \mathbb{R}^k$. A **statistic** is a random variable of the form $Y := f(X_1, \ldots, X_n)$. The distribution of Y is called a **sampling distribution**.

Example 4.3. The **sample mean** of a random sample $X_1, \ldots, X_n$ of size n, denoted $\overline{X}$, is the following statistic:

$$\overline{X} := \frac{1}{n} \sum_{i=1}^n X_i.$$

Example 4.4. Let $n > 1$. The **sample standard deviation** of a random sample $X_1, \ldots, X_n$ of size n, denoted S, is the following statistic:

$$S := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2}.$$

The **sample variance** of a random sample $X_1, \ldots, X_n$ of size n is S^2.

From the usual definition of the variance (for the uniform distribution on the integers $\{1, \ldots, n\}$), it might seem sensible to divide by n above instead of $n-1$. The second part of the following exercise attempts to explain why dividing by $n-1$ is sensible.

Exercise 4.5. Let $n \geq 2$ be an integer. Let $X_1, \ldots, X_n$ be a random sample of size n. Assume that $\mu := \mathbf{E}X \in \mathbb{R}$ and $\sigma := \sqrt{\mathrm{var}(X)} < \infty$. Let $\overline{X}$ be the sample mean and let S be the sample standard deviation of the random sample. Show the following

- $\mathrm{Var}(\overline{X}) = \sigma^2/n$.
- $\mathbf{E}S^2 = \sigma^2$.

If we divided by n instead of $n-1$ in the definition of S, then the second part of the above exercise would not hold. Since $\mathbf{E}S^2$ agrees with the variance of X, we say that S^2 is unbiased. We will discuss this concept more in Section 6.

Exercise 4.6. Let $X \colon \Omega \to \mathbb{R}$ be a random variable with $\mathbf{E}X^2 < \infty$. Show that the quantity $\mathbf{E}(X-t)^2$ is minimized for $t \in \mathbb{R}$ uniquely when $t = \mathbf{E}X$.

4.1. **Sampling from the Normal.** The Central Limit Theorem implies that the combination of a large number of independent identically distributed random actions results in a Gaussian distribution. For this reason, one can often (but not always) assume that sampling from a large population is sampling from the normal distribution with unknown mean and variance. Since this Gaussian assumption is so common, we discuss properties of sampling from the normal in this section.

Proposition 4.7. *Let $n \geq 2$ be an integer. Let $X_1, \ldots, X_n$ be a random sample from the Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Let $\overline{X}$ be the sample mean and let S be the sample standard deviation.*

- *$\overline{X}$ and S are independent random variables.*
- *$\overline{X}$ is a Gaussian random variable with mean μ and variance σ^2/n.*
- *$(n-1)S^2/\sigma^2$ is a chi-squared distributed random variable with $n-1$ degrees of freedom.*

Proof. By replacing $X_1, \ldots, X_n$ with $X_1 - \mu, \ldots, X_n - \mu$, it suffices to assume that $\mu = 0$ in the proof. It further suffices to assume $\sigma = 1$ by dividing all the random variables by σ. To prove the first item, we first note that the random variable $\overline{X}$ is independent the collection of random variables $X_1 - \overline{X}, \ldots, X_n - \overline{X}$. This follows from Exercise 3.16, since the vector $(1, \ldots, 1) \in \mathbb{R}^n$ is orthogonal to any vector in the span of

$$(1,0,0,\ldots) - \frac{1}{n}(1,\ldots,1), \quad \ldots \quad , (0,\ldots,0,1) - \frac{1}{n}(1,\ldots,1).$$

(We are not asserting that the random variables $\overline{X}, X_1 - \overline{X}, \ldots, X_n - \overline{X}$ are all independent; in fact this is false by Exercise 3.16 since the vectors $(1,0,0,\ldots) - \frac{1}{n}(1,\ldots,1), (0,0,\ldots,0,1) - \frac{1}{n}(1,\ldots,1)$ are not orthogonal in $\mathbb{R}^n$.) The first item follows, since S is a function of $X_1 - \overline{X}, \ldots, X_n - \overline{X}$, so S is independent of $\overline{X}$.

The second item follows from Proposition 1.45, Example 1.108 and Exercise 1.58.

We now prove the third item. Let $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ and let $S_n^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. In the case $n = 2$, we have $S_2^2 = \frac{1}{4}(X_1 - X_2)^2 + \frac{1}{4}(X_2 - X_1)^2 = \frac{1}{2}(X_1 - X_2)^2$. From Example 1.108, $\frac{1}{\sqrt{2}}(X_1 - X_2)$ is a mean zero Gaussian random variable with variance 1. So, S_2^2 is a chi-squared distributed random variable by Definition 1.33 with one degree of freedom. That is, the third item of this proposition holds when $n = 2$.

We now induct on n. From Lemma 4.8,

$$nS_{n+1}^2 = (n-1)S_n^2 + \frac{n}{n+1}(X_{n+1} - \overline{X}_n)^2, \qquad \forall\, n \geq 2.$$

From the first item, S_n is independent of $\overline{X}_n$. Also, X_{n+1} is independent of S_n by Proposition 1.61, since S_n is a function of $X_1, \ldots, X_n$, the latter being independent of X_{n+1}. In summary, S_n is independent of $(X_{n+1} - \overline{X}_n)^2$. By the inductive hypothesis, $(n-1)S_n^2$ is a chi-squared distributed random variable with $n-1$ degrees of freedom. From Example 1.108 $X_{n+1} - \overline{X}_n$ is a Gaussian random variable with mean zero and variance $1+1/n$, so that $\sqrt{n/(n+1)}(X_{n+1} - \overline{X}_n)$ is a mean zero Gaussian with variance 1. Definition 1.33 then implies that nS_{n+1} is a chi-squared random variable with n degrees of freedom, completing the inductive step. □

Lemma 4.8. *Let $X_1, X_2, \ldots$ be random variables. For any $n \geq 2$, let $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ and let $S_n^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Then*

$$nS_{n+1}^2 - (n-1)S_n^2 = \frac{n}{n+1}(X_{n+1} - \overline{X}_n)^2, \qquad \forall\, n \geq 2.$$

Proof.

$$\begin{aligned}
nS_{n+1}^2 - (n-1)S_n^2 &= \sum_{i=1}^{n+1}(X_i - \overline{X}_{n+1})^2 - \sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \\
&= (X_{n+1} - \overline{X}_{n+1})^2 + \sum_{i=1}^{n}(\overline{X}_{n+1} - \overline{X}_n)(-2X_i + \overline{X}_{n+1} + \overline{X}_n) \\
&= (X_{n+1} - \overline{X}_{n+1})^2 + (\overline{X}_{n+1} - \overline{X}_n)\sum_{i=1}^{n}(-2X_i + \overline{X}_{n+1} + \overline{X}_n) \\
&= (X_{n+1} - \overline{X}_{n+1})^2 + (\overline{X}_{n+1} - \overline{X}_n)n(-2\overline{X}_n + \overline{X}_{n+1} + \overline{X}_n) \\
&= (X_{n+1}(1 - 1/(n+1)) - \frac{n}{n+1}\overline{X}_n)^2 + n(\overline{X}_{n+1} - \overline{X}_n)^2 \\
&= \frac{n^2}{(n+1)^2}(X_{n+1} - \overline{X}_n)^2 + n\Big(\frac{X_{n+1}}{n+1} + (\frac{1}{n+1} - \frac{1}{n})\sum_{i=1}^{n} X_i\Big)^2 \\
&= \frac{n^2}{(n+1)^2}(X_{n+1} - \overline{X}_n)^2 + n\Big(\frac{X_{n+1}}{n+1} - \frac{1}{n(n+1)}\sum_{i=1}^{n} X_i\Big)^2 \\
&= \frac{n^2}{(n+1)^2}(X_{n+1} - \overline{X}_n)^2 + n\Big(\frac{X_{n+1}}{n+1} - \frac{1}{n+1}\overline{X}_n\Big)^2 \\
&= \frac{n^2}{(n+1)^2}(X_{n+1} - \overline{X}_n)^2 + \frac{n}{(n+1)^2}(X_{n+1} - \overline{X}_n)^2 = \frac{n}{n+1}(X_{n+1} - \overline{X}_n)^2.
\end{aligned}$$

□

If $X_1, X_2, \ldots$ are a random sample from a Gaussian random variable with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$, then Example 1.108 implies that

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

is a Gaussian random variable with mean zero and variance one. If the mean and standard deviation are unknown, then it might be difficult to find either μ or σ by looking at this

quantity for different values of μ and σ. However, if we substitute the sample variance S for σ and examine instead

$$\frac{\overline{X}-\mu}{S/\sqrt{n}},$$

then there is only one unknown parameter μ appearing in this expression. So, if we insert different values of μ into $\frac{\overline{X}-\mu}{S/\sqrt{n}}$, we might be able to determine the unknown mean μ, if we knew the distribution of $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ for fixed μ. This distribution is given by the following proposition.

Proposition 4.9. *Let X be a standard Gaussian random variable. Let Y be a chi squared random variable with p degrees of freedom. Assume that X and Y are independent. Then $X/\sqrt{Y/p}$ has the following density, known as* ***Student's t-distribution*** *with p degrees of freedom:*

$$f_{X/(\sqrt{Y/p})}(t) := \frac{\Gamma(\frac{p+1}{2})}{\sqrt{p}\sqrt{\pi}\Gamma(p/2)}\Big(1+\frac{t^2}{p}\Big)^{-\frac{p+1}{2}}, \qquad \forall\, t\in\mathbb{R}.$$

Remark 4.10. If $X_1,\dots,X_{n+1}$ is a random sample from a Gaussian random variable with mean $\mu\in\mathbb{R}$ and standard deviation $\sigma>0$, then $(n-1)^{-1/2}(\overline{X}-\mu)/(S/\sqrt{n})$ has Student's t-distribution with $n-1$ degrees of freedom, since $\sqrt{n}(\overline{X}-\mu)$ has mean zero and variance one, and dividing the top and bottom by σ reduces to the case treated in the proposition (using also independence of $\overline{X}$ and S by Proposition 4.7).

Proof. First, let $Z := \sqrt{Y/p}$. We find the density of Z as follows. Let $t>0$. Then

$$\begin{aligned} f_Z(y) &= \frac{d}{dy}\mathbf{P}(Z\le y) = \frac{d}{dy}\mathbf{P}(Y\le y^2p) = \frac{d}{dy}\int_0^{y^2p}\frac{x^{(p/2)-1}e^{-x/2}}{2^{p/2}\Gamma(p/2)}dx \\ &= 2ypp^{(p/2)-1}y^{p-2}e^{-y^2p/2}\frac{1}{2^{p/2}\Gamma(p/2)} = p^{p/2}y^{p-1}e^{-y^2p/2}\frac{1}{2^{(p/2)-1}\Gamma(p/2)}. \end{aligned}$$

Let $\phi\colon\mathbb{R}^2\to\mathbb{R}^2$ so that $\phi^{-1}(x,y)=(y,x/y)$, $\phi(a,b)=(ab,a)$, $|\mathrm{Jac}\phi(a,b)| = \left|\det\begin{pmatrix} b & a\\ 1 & 0\end{pmatrix}\right| = |a|$, for all $(x,y),(a,b)\in\mathbb{R}^2$. By the Change of Variables formula, for any $U\subseteq\mathbb{R}^2$,

$$\iint_{\phi(U)} f(x,y)dxdy = \iint_U f(\phi(a,b))\,|\mathrm{Jac}\phi(a,b)|\,dadb.$$

Let $t>0$. Then by the definition of the joint distribution, and independence of X,Z,

$$\begin{aligned} \mathbf{P}(\frac{X}{Z}\le t) = \mathbf{P}(X\le tZ) &= \int_{\{(x,y)\in\mathbb{R}^2\colon x\le ty,\, y>0\}} f_X(x)f_Z(y)dxdy \\ &= \int_{\{(a,b)\in\mathbb{R}^2\colon b\le t,\, a>0\}} |a|\, f_X(ab)f_Z(a)dadb = \int_{b=-\infty}^{b=t}\int_{a=0}^{\infty}|a|\, f_X(ab)f_Z(a)dadb. \end{aligned}$$

So, taking the derivative in t,applying the Fundamental Theorem of Calculus, and using the change of variables $x = a^2$ so that $da = \frac{1}{2\sqrt{x}}dx$,

$$f_{X/Z}(t) = \int_0^\infty |a|\, f_X(at) f_Z(a) da = \frac{p^{p/2}}{\sqrt{2\pi}} \int_0^\infty a e^{-a^2t^2/2} a^{p-1} e^{-a^2 p/2} \frac{1}{2^{(p/2)-1}\Gamma(p/2)} da$$
$$= \frac{p^{p/2}}{\sqrt{2\pi}2^{(p/2)-1}\Gamma(p/2)} \int_0^\infty a^p e^{-(p+t^2)a^2/2} da = \frac{p^{p/2}}{\sqrt{2\pi}2^{p/2}\Gamma(p/2)} \int_0^\infty x^{(p/2)-1/2} e^{-(p+t^2)x/2} dx.$$

From Definition 1.33, the integrand is the density of a gamma distributed random variable with parameters α, β where $\alpha - 1 = (p/2) - 1/2$ and $\beta = 2/(p+t^2)$; so that if we divide and multiply by $\beta^\alpha \Gamma(\alpha)$, we have

$$f_{X/Z}(t) = \frac{p^{p/2}}{\sqrt{2\pi}2^{p/2}\Gamma(p/2)} \beta^\alpha \Gamma(\alpha) \cdot (1) = \frac{p^{p/2}\Gamma(\frac{p+1}{2})}{\sqrt{2\pi}2^{p/2}\Gamma(p/2)} \Big(\frac{2}{p+t^2}\Big)^{\frac{p+1}{2}}$$
$$= \frac{p^{p/2}\Gamma(\frac{p+1}{2})}{\sqrt{2\pi}2^{p/2}\Gamma(p/2)} 2^{(p+1)/2} p^{-(p+1)/2} \Big(1 + \frac{t^2}{p}\Big)^{-\frac{p+1}{2}} = \frac{\Gamma(\frac{p+1}{2})}{\sqrt{p}\sqrt{\pi}\Gamma(p/2)} \Big(1 + \frac{t^2}{p}\Big)^{-\frac{p+1}{2}}.$$

□

Remark 4.11. The definition of Student's t distribution looks not quite right in the Casella and Berger book.

Exercise 4.12. Let X be a chi squared random variables with p degrees of freedom. Let Y be a chi squared random variable with q degrees of freedom. Assume that X and Y are independent. Show that $(X/p)/(Y/q)$ has the following density, known as **Snedecor's f-distribution** with p and q degrees of freedom

$$f_{(X/p)/(Y/q)}(t) := \frac{t^{(p/2)-1}(p/q)^{p/2}\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)} \Big(1 + t(p/q)\Big)^{-(p+q)/2}, \qquad \forall\, t > 0.$$

Exercise 4.13 (Order Statistics). Let $X : \Omega \to \mathbb{R}$ be a random variable. Let $X_1, \ldots, X_n$ be a random sample of size n from X. Define $X_{(1)} := \min_{1 \le i \le n} X_i$, and for any $2 \le i \le n$, inductively define

$$X_i := \min\Big\{\{X_1, \ldots, X_n\} \smallsetminus \{X_{(1)}, \ldots, X_{(i-1)}\}\Big\},$$

so that

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)} = \max_{1 \le i \le n} X_i.$$

The random variables $X_{(1)}, \ldots, X_{(n)}$ are called the **order statistics** of $X_1, \ldots, X_n$.

- Suppose X is a discrete random variable and we can order the values that X takes as $x_1 < x_2 < \cdots$. For any $i \ge 1$, define $p_i := \mathbf{P}(X \le x_i)$. Show that, for any $1 \le i, j \le n$,

$$\mathbf{P}(X_{(j)} \le x_i) = \sum_{k=j}^{n} \binom{n}{k} p_i^k (1 - p_i)^{n-k}.$$

(Hint: Let Y be the number of indices $1 \le j \le n$ such that $X_j \le x_i$. Then Y is a binomial random variable with parameters n and p_i.)

You don't have to show it, but if X is a continuous random variable with density f_X and cumulative distribution function F_X, then for any $1 \le j \le n$, $F_{X_{(j)}}$ has density

$$f_{X_{(j)}}(x) := \frac{n!}{(j-1)!(n-j)!} f_X(x)(F_X(x))^{j-1}(1-F_X(x))^{n-j}, \qquad \forall\, x \in \mathbb{R}.$$

(This follows by differentiating the above identity for the cumulative distribution function.)

- Let X be a random variable uniformly distributed in $[0,1]$. For any $1 \le j \le n$, show that $X_{(j)}$ is a beta distributed random variable with parameters j and $n-j$. Conclude that (as you might anticipate)

$$\mathbf{E}X_{(j)} = \frac{j}{n+1}.$$

- Let $a, b \in \mathbb{R}$ with $a < b$. Let U be the number of indices $1 \le j \le n$ such that $X_j \le a$. Let V be the number of indices $1 \le j \le n$ such that $a < X_j \le b$. Show that the vector $(U, V, n-U-V)$ is a multinomial random variable, so that for any nonnegative integers u, v with $u+v \le n$, we have

$$\begin{aligned}&\mathbf{P}(U=u, V=v, n-U-V=n-u-v)\\ &\quad= \frac{n!}{u!v!(n-u-v)!} F_X(a)^u (F_X(b)-F_X(a))^v (1-F_X(b))^{n-u-v}.\end{aligned}$$

Consequently, for any $1 \le i, j \le n$,

$$\mathbf{P}(X_{(i)} \le a, X_{(j)} \le b) = \mathbf{P}(U \ge i, U+V \ge j) = \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \mathbf{P}(U=k, V=m) + \mathbf{P}(U \ge j).$$

So, it is possible to write an explicit formula for the joint distribution of $X_{(i)}$ and $X_{(j)}$ (but you don't have to write it yourself).

4.2. **The Delta Method.** From Examples 4.3 and 4.4 and Exercise 4.5, the sample mean and sample variance give good estimates for the mean and variance of random samples. More generally, we might want an estimate for a function of the mean or a function of the variance. Such an estimate is provided by the following version of the Central Limit Theorem.

Theorem 4.14 (Delta Method). *Let $\theta \in \mathbb{R}$. Let $Y_1, Y_2, \ldots$ be random variables such that $\sqrt{n}(Y_n - \theta)$ converges in distribution to a mean zero Gaussian random variable with variance $\sigma^2 > 0$. Let $f\colon \mathbb{R} \to \mathbb{R}$. Assume that $f'(\theta)$ exists. Then*

$$\sqrt{n}(f(Y_n) - f(\theta))$$

converges in distribution to a mean zero Gaussian with variance $\sigma^2(f'(\theta))^2$ as $n \to \infty$.

Proof. Since $f'(\theta)$ exists, $\lim_{y\to\theta} \frac{f(y)-f(\theta)}{y-\theta}$ exists. That is, there exists $h\colon \mathbb{R} \to \mathbb{R}$ such that $\lim_{z\to 0} \frac{h(z)}{z} = 0$ and, for all $y \in \mathbb{R}$,

$$f(y) = f(\theta) + f'(\theta)(y-\theta) + h(y-\theta).$$

In particular,

$$\sqrt{n}[f(Y_n) - f(\theta)] = f'(\theta)\sqrt{n}(Y_n - \theta) + \sqrt{n}h(Y_n - \theta). \qquad (*)$$

By assumption, $\forall\ s,t>0$, $\lim_{n\to\infty}\mathbf{P}(|Y_n-\theta|>st/\sqrt{n})=2\int_{st}^{\infty}e^{-y^2/[2\sigma^2]}\frac{dy}{\sigma\sqrt{2\pi}}$. So, $\forall\ n\geq 1$,

$$\begin{aligned}\mathbf{P}(\sqrt{n}\,|h(Y_n-\theta)|>t)&=\mathbf{P}(\sqrt{n}\,|h(Y_n-\theta)|>t,\ |Y_n-\theta|>st/\sqrt{n})\\&\quad+\mathbf{P}(\sqrt{n}\,|h(Y_n-\theta)|>t,\ |Y_n-\theta|\leq st/\sqrt{n})\\&\leq\mathbf{P}(|Y_n-\theta|>st/\sqrt{n})+\mathbf{P}(\sqrt{n}\,|h(Y_n-\theta)|>t,\ |Y_n-\theta|\leq st/\sqrt{n}).\end{aligned}$$

As $n\to\infty$, the first term converges to $2\int_{st}^{\infty}e^{-\frac{y^2}{2\sigma^2}}\frac{dy}{\sigma\sqrt{2\pi}}$, and the second term goes to zero since $\lim_{z\to 0}(h(z)/z)=0$. So, for any $s,t>0$, $\lim_{n\to\infty}\mathbf{P}(\sqrt{n}\,|h(Y_n-\theta)|>t)\leq 2\int_{st}^{\infty}e^{-\frac{y^2}{2\sigma^2}}\frac{dy}{\sigma\sqrt{2\pi}}$. Since this holds for any $s>0$, we can let $s\to\infty$ to get $\lim_{n\to\infty}\mathbf{P}(\sqrt{n}\,|h(Y_n-\theta)|>t)=0$. That is, $\sqrt{n}h(Y_n-\theta)$ converges in probability to zero as $n\to\infty$. So, by Proposition 2.36 and $(*)$, $\sqrt{n}[f(Y_n)-f(\theta)]$ converges in distribution to a mean zero Gaussian with variance $\sigma^2(f'(\theta))^2$. □

Example 4.15. Suppose $\overline{X}_n$ is the sample mean for a random sample $X_1,\ldots,X_n$ of size n and $0<\text{var}(X_1)<\infty$. Let $\mu:=\mathbf{E}X_1\neq 0$. From the Central Limit Theorem 2.13, $\sqrt{n}(\overline{X}_n-\mu)$ converges in distribution to a mean zero Gaussian with variance $\sigma^2:=\text{var}(X_1)$. So, if we use $f(x):=1/x$ for any $x\neq 0$, the random variable $\sqrt{n}(f(\overline{X}_n)-\frac{1}{\mu})$ converges in distribution to a mean zero Gaussian with variance $\sigma^2(f'(\mu))^2=\sigma^2\mu^{-4}$ as $n\to\infty$.

From Exercises 2.6 and 2.7, this does *not* imply that the variance of $\sqrt{n}(f(\overline{X}_n)-1/\mu)$ converges. However, if we assume there exists $\varepsilon,c>0$ such that $\mathbf{E}\left|\sqrt{n}(f(\overline{X}_n)-\frac{1}{\mu})\right|^{2+\varepsilon}\leq c$ for all $n\geq 1$, then we can conclude that

$$\lim_{n\to\infty}\text{var}(\sqrt{n}(f(\overline{X}_n)-\frac{1}{\mu}))=\sigma^2(f'(\mu))^2$$

by Theorem 4.16 below with $X_n':=(f(\overline{X}_n)-\frac{1}{\mu})^2$ for all $n\geq 1$.

So, we can say that $1/\overline{X}_n$ has expected value near $1/\mu$ variance near $n^{-1}\sigma^2\mu^{-4}$, when n is large.

Theorem 4.16 (Convergence Theorem with Bounded Moment). *Let $X_1,X_2,\ldots$ be random variables that converge in distribution to a random variable X. Assume $\exists\ 0<\varepsilon,c<\infty$ such that $\mathbf{E}\,|X_n|^{1+\varepsilon}\leq c$, $\forall\ n\geq 1$. Then*

$$\mathbf{E}X=\lim_{n\to\infty}\mathbf{E}X_n.$$

For a proof, see my Graduate Probability Notes (Theorem 1.59 together with Exercise 3.8(iii).)

In the case that $f'(\theta)=0$ in the Delta Method, we can instead use a second order Taylor expansion as follows.

Theorem 4.17 (Second Order Delta Method). *Let $\theta\in\mathbb{R}$. Let $Y_1,Y_2,\ldots$ be random variables such that $\sqrt{n}(Y_n-\theta)$ converges in distribution to a mean zero Gaussian random variable with variance $\sigma^2>0$. Let $f\colon\mathbb{R}\to\mathbb{R}$. Assume that $f'(\theta)=0$, $f''(\theta)$ exists and is nonzero. Then*

$$n(f(Y_n)-f(\theta))$$

converges in distribution to a chi squared random variable with one degree of freedom, multiplied by $\sigma^2\frac{1}{2}f''(\theta)$ as $n\to\infty$.

Proof. Since $f'(\theta) = 0$, there exists $g\colon \mathbb{R} \to \mathbb{R}$ such that $\lim_{z\to 0} g(z) = 0$ and, for all $y \in \mathbb{R}$,

$$f(y) = f(\theta) + (y-\theta)g(y-\theta).$$

Since $f''(\theta)$ exists, the following limit exists

$$\lim_{s\to 0} \frac{f(\theta+2s) + f(\theta) - 2f(\theta+s)}{s^2} = \lim_{s\to 0} \frac{2sg(2s) - 2sg(s)}{s^2} = \lim_{s\to 0} 2\frac{g(2s)-g(s)}{s} = 2g'(0).$$

Since $g'(0)$ exists, there exists $r\colon \mathbb{R} \to \mathbb{R}$ such that $\lim_{z\to 0} \frac{r(z)}{z} = 0$ and, for all $y \in \mathbb{R}$,

$$g(y) = g(0) + g'(0)y + r(y).$$

Since $g'(0)$ exists, g is continuous at 0, so $g(0) = \lim_{z\to 0} g(z) = 0$. Combining the above, for all $y \in \mathbb{R}$,

$$\begin{aligned} f(y) &= f(\theta) + (y-\theta)g(0) + (y-\theta)^2 g'(0) + (y-\theta)r(y-\theta) \\ &= f(\theta) + (y-\theta)^2 \frac{1}{2} f''(\theta) + (y-\theta)r(y-\theta). \end{aligned}$$

Let $h(y) := yr(y)$ for all $y \in \mathbb{R}$. Then $\lim_{y\to 0} \frac{h(y)}{y^2} = \lim_{y\to 0} \frac{r(y)}{y} = 0$. Also,

$$n[f(Y_n) - f(\theta)] = \frac{1}{2} f''(\theta) n(Y_n - \theta) + nh(Y_n - \theta). \qquad (*)$$

By assumption, for all $s, t > 0$, $\lim_{n\to\infty} \mathbf{P}(|Y_n - \theta| > st/\sqrt{n}) = 2\int_{st}^{\infty} e^{-\frac{y^2}{2\sigma^2}} \frac{dy}{\sigma\sqrt{2\pi}}$. So, $\forall\, n \geq 1$,

$$\begin{aligned} \mathbf{P}(n\,|h(Y_n-\theta)| > t) &= \mathbf{P}(n\,|h(Y_n-\theta)| > t,\ |Y_n - \theta| > st/\sqrt{n}) \\ &\quad + \mathbf{P}(n\,|h(Y_n-\theta)| > t,\ |Y_n - \theta| \leq st/\sqrt{n}) \\ &\leq \mathbf{P}(|Y_n - \theta| > st/\sqrt{n}) + \mathbf{P}(n\,|h(Y_n-\theta)| > t,\ |Y_n - \theta| \leq st/\sqrt{n}). \end{aligned}$$

As $n \to \infty$, the first term goes to $2\int_{st}^{\infty} e^{-y^2/[2\sigma^2]} \frac{dy}{\sigma\sqrt{2\pi}}$, and the second term goes to zero since $\lim_{z\to 0}(h(z)/z^2) = 0$. So, for any $s, t > 0$, $\lim_{n\to\infty} \mathbf{P}(n\,|h(Y_n-\theta)| > t) \leq 2\int_{st}^{\infty} e^{-\frac{y^2}{2\sigma^2}} \frac{dy}{\sigma\sqrt{2\pi}}$. Since this holds for any $s > 0$, we can let $s \to \infty$ to get $\lim_{n\to\infty} \mathbf{P}(n\,|h(Y_n-\theta)| > t) = 0$. That is, $nh(Y_n - \theta)$ converges in probability to zero as $n \to \infty$. So, by Proposition 2.36 and $(*)$, $n[f(Y_n) - f(\theta)]$ converges in distribution to a chi squared random variable with one degree of freedom, multiplied by $\sigma^2 f''(\theta)/2$. □

Let $m > 2$ be an integer. Theorem 4.17 generalizes to: if $f'(\theta) = \cdots = f^{(m-1)}(\theta) = 0$, if $f^{(m)}(\theta)$ exists and is nonzero, then as $n \to \infty$,

$$n^{m/2}(f(Y_n) - f(\theta))$$

converges in distribution to the distribution of a standard Gaussian to the m^{th} power, multiplied by $\sigma^m \frac{1}{m!} f^{(m)}(\theta)$.

4.3. **Simulation of Random Variables.** In practice we often want to simulate random variables on a computer. The sampling of random variables on a computer is also called **Monte Carlo simulation**. In this section, we assume that a computer can simulate any number of independent random variable that are uniformly distributed in $(0, 1)$. From this assumption, we will try to transform that random variable into other ones.

There are some caveats to our assumption that we can sample from the uniform distribution on $(0, 1)$.

(1) Computers cannot deal with arbitrary real numbers. The most common number system used on computers is instead **double precision floating point arithmetic**. This number system includes zero and any number of the form

$$\pm(1.a_1a_2\cdots a_{52})\cdot 2^{b_1\cdots b_{11}-1023},$$

where $a_1,\ldots,a_{52},b_1,\ldots,b_{11}\in\{0,1\}$ are binary digits, and $b_1,\ldots,b_{11}$ are not all 0 and not all 1. Consequently, a computer can at best simulate a number that is drawn randomly from the 2^{64} numbers of this form. Put another way, every random variable simulated on a computer is automatically discrete.

(2) A computer cannot produce a truly random quantity. When we repeatedly sample from a random variable on a computer, the computer uses a deterministic process to produce a sequence of numbers that behaves as if it were random. For this reason, random number generators on computers are said to produce **pseudorandom** outputs. There are a various random number generating algorithms available.

We can verify that a random number generator behaves "as if it were random" by checking for its agreement with the Law of Large Number and Central Limit Theorem.

Exercise 4.18. Using Matlab (or any other mathematical system on a computer), verify that its random number generator agrees with the law of large numbers and central limit theorem. For example, average 10^7 samples from the uniform distribution on $[0,1]$ and check how close the sample average is to $1/2$. Then, make a histogram of 10^7 samples from the uniform distribution on $[0,1]$ and check how close the histogram is to a Gaussian.

Example 4.19 (**Discrete Random Variables**)**.** If we want to simulate a random variable that is uniformly distributed in $\{1,2,3\}$, and if U is uniform on $(0,1)$, we define

$$X(U):=\begin{cases}1 & \text{if } U<1/3\\ 2 & \text{if } 1/3\le U<2/3\\ 3 & \text{if } 2/3\le U.\end{cases}$$

Then $X(U)$ is uniformly distributed in $\{1,2,3\}$.

More generally, if we want to simulate a random variable taking values $x_1,\ldots,x_n\in\mathbb{R}$ with probabilities $p_1,\ldots,p_n>0$ such that $p_1+\cdots+p_n=1$, we define $p_0:=0$ and we define $X(U)$ so that

$$X(U):=x_i \qquad \text{if} \quad p_1+\cdots+p_{i-1}\le U<p_1+\cdots+p_i \;\forall 1\le i\le n.$$

Then $\mathbf{P}(X(U)=x_i)=p_i$ for all $1\le i\le n$, as desired.

More generally, if $X\colon\Omega\to\mathbb{R}$ is an arbitrary random variable with cumulative distribution function $F\colon\mathbb{R}\to[0,1]$, then the function F^{-1} (if it exists) is a random variable on $[0,1]$ with the uniform probability law on $(0,1)$ that is equal in distribution to X, since

$$\mathbf{P}(s\in[0,1]\colon F^{-1}(s)\le t)=\mathbf{P}(s\in[0,1]\colon F(t)>s)\overset{(*)}{=}F(t)=\mathbf{P}(\omega\in\Omega\colon X(\omega)\le t).$$

Here $(*)$ used the definition of the uniform probability law on $(0,1)$. In general, F^{-1} may not exist, but we can still construct a generalized inverse of F and obtain the same conclusion as follows.

Exercise 4.20. Let $X\colon \Omega \to \mathbb{R}$ be a random variable on a sample space Ω equipped with a probability law $\mathbf{P}$. For any $t \in \mathbb{R}$ let $F(t) := \mathbf{P}(X \leq t)$. For any $s \in (0,1)$ define

$$Y(s) := \sup\{t \in \mathbb{R}\colon F(t) < s\}.$$

Then Y is a random variable on $(0,1)$ with the uniform probability law on $(0,1)$. Show that X and Y are equal in distribution. That is, $\mathbf{P}(Y \leq t) = F(t)$ for all $t \in \mathbb{R}$.

Exercise 4.20 then suggest the following method for simulating a random variable on a computer.

Algorithm 4.21 (**Sampling a Random Variable**)**.** Let $X\colon \Omega \to \mathbb{R}$ be a random variable. Let $\mathbf{P}$ be a probability law on Ω. For any $t \in \mathbb{R}$, let $F(t) := \mathbf{P}(X \leq t)$. Let U be a random variable uniformly distributed in $(0,1)$. For any $s \in (0,1)$, let

$$Y(s) := \sup\{t \in \mathbb{R}\colon F(t) < s\}.$$

To sample X on a computer, sample $Y(U)$.

Example 4.22. Let X be an exponential random variable with parameter 1, so that for any $t > 0$, $\mathbf{P}(X \leq t) = \int_0^t e^{-x}dx = 1 - e^{-t} =: F(t)$. Then $F^{-1}(s) = -\log(1-s)$ for any $0 < s < 1$, since $F(F^{-1}(s)) = s$. By Exercise 4.20, F^{-1} is an exponential random variable with parameter 1 if $\mathbf{P}$ is the uniform probability law on $(0,1)$. Or by Algorithm 4.21, $F^{-1}(U) = -\log(1-U)$ is an exponential random variable with parameter 1.

When an explicit formula can be given for Y in Algorithm 4.21, the random variable can be simulated efficiently. However, if Y cannot be accurately or efficiently computed, Algorithm 4.21 may not be a sensible way to simulate a random variable. For example, consider a standard Gaussian random variable. The inverse of its cumulative distribution function cannot be described using elementary formulas. Here are some possible ways to simulate a standard Gaussian.

- Approximate the inverse cumulative distribution function and apply Algorithm 4.21. The quality of the approximation then correspond to the quality of the simulation.
- Sample many independent uniform random variables $U_1, \ldots, U_n$ in $(0,1)$. Form the sum $\frac{U_1+\cdots+U_n-n/2}{n\sqrt{1/12}}$. By the Central Limit Theorem 2.13, this random variable is close to a standard Gaussian. In fact, explicit error bounds can be given by Theorem 2.30. Moreover, if we perform this same procedure where $U_1, \ldots, U_n$ are i.i.d. and the first k moments of U_1 agree with the first k moments of a standard Gaussian, the error in Theorem 2.30 will be a constant times $n^{-(k-1)/2}$. (This follows from the Edgeworth expansion, an asymptotic expansion for the error in the Central Limit Theorem.) However, if we only want a few samples from the Gaussian, this procedure is very inefficient, since it requires many samples from other random variables.

Perhaps the best way to simulate a standard Gaussian random variable is the Box-Mueller algorithm.

Exercise 4.23 (**Box-Muller Algorithm**)**.** Let U_1, U_2 be independent random variables uniformly distributed in $(0,1)$. Define

$$R := \sqrt{-2\log U_1}, \qquad \Psi := 2\pi U_2.$$
$$X := R\cos\Psi, \qquad Y := R\sin\Psi.$$

Show that X, Y are independent standard Gaussian random variables. So, we can simulate any number of independent standard Gaussian random variables with this procedure.

Now, let $\{a_{ij}\}_{1\le i,j\le n}$ be an $n \times n$ symmetric positive semidefinite matrix. That is, for any $v \in \mathbb{R}^n$, we have

$$v^T a v = \sum_{i,j=1}^{n} v_i v_j a_{ij} \ge 0.$$

We can simulate a Gaussian random vector with any such covariance matrix $\{a_{ij}\}_{1\le i,j\le n}$ using the following procedure.

- Let $X = (X_1, \dots, X_n)$ be a vector of i.i.d. standard Gaussian random variables (which can be sampled using the Box-Muller algorithm above).
- Write the matrix a in its Cholesky decomposition $a = rr^*$, where r is an $n \times n$ real matrix. (This decomposition can be computed efficiently with about n^3 arithmetic operations.)
- Let $e^{(1)}, \dots, e^{(n)}$ be the rows of r. For any $1 \le i \le n$, define

$$Z_i := \langle X, e^{(i)} \rangle.$$

Show that $Z := (Z_1, \dots, Z_n)$ is a mean zero Gaussian random vector whose covariance matrix is $\{a_{ij}\}_{1\le i,j\le n}$, so that

$$\mathbf{E}(Z_i Z_j) = a_{ij}, \qquad \forall\, 1 \le i, j \le n.$$

4.4. **Additional Comments.** The assumption that astronomical data sampling error arose from sampling from the normal distribution was common in the early 1800s, and Quetelet was one of the first of that period to apply the normal assumption to other scientific fields. In the late 1700s, Laplace's applications of statistics were also eclectic. The Delta Method was known in the early 1800s, though it was not precisely described until the early 1900s.

Theorem 4.24 (**Multivariate Delta Method**). *Let $\Theta \subseteq \mathbb{R}^m$ be an open set. Let $\theta \in \Theta$. Let $Y_1, Y_2, \dots \in \mathbb{R}^m$ be random variables such that $\sqrt{n}(Y_n - \theta)$ converges in distribution to a mean zero Gaussian random vector with covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$. Let $f\colon \mathbb{R}^m \to \mathbb{R}^m$. Assume that f is differentiable at θ. Let $Dg(\theta)$ denote the matrix of first order partial derivatives of g. Then as $n \to \infty$,*

$$\sqrt{n}(f(Y_n) - f(\theta))$$

converges in distribution to a mean zero Gaussian vector with covariance matrix

$$(Df(\theta))^T \Sigma Df(\theta).$$

5. Data Reduction

Suppose we have some data and an exponential family. We would like to find the parameter θ among the exponential family that fits the data well. One way to achieve this goal is to look for a sufficient statistic.

5.1. **Sufficient Statistics.**

Definition 5.1 (**Sufficient Statistic**). Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a distribution f where $f \in \{f_\theta \colon \theta \in \Theta\}$ is a family of PDFs or PMFs (such as an exponential family). Let $t \colon \mathbb{R}^n \to \mathbb{R}^k$, so that $Y := t(X_1, \ldots, X_n)$ is a statistic. We say that Y is a **sufficient statistic** for θ if, for every $y \in \mathbb{R}^k$ and for every $\theta \in \Theta$, the conditional distribution of $(X_1, \ldots, X_n)$ given $Y = y$ (with respect to probabilities given by f_θ) does not depend on θ. That is, Y provides sufficient information to determine θ from $X_1, \ldots, X_n$.

Note that any invertible function of a sufficient statistic is sufficient.

Also, the term "sufficient" is a bit misleading. A sufficient statistic does not contain sufficient information to *exactly* determine the parameter θ. As we will see in the next example, the sample mean is a sufficient statistic for the Bernoulli distribution, but this does not mean that we can exactly determine the unknown parameter of the Bernoulli. Being a sufficient statistic essentially means that we can make the best possible guess for the unknown parameter using the sufficient statistic.

Example 5.2. Let $X_1, \ldots, X_n$ be a random sample of size n from a Bernoulli distribution with parameter $0 < \theta < 1$. We claim that $Y := X_1 + \cdots + X_n$ is a sufficient statistic for θ. Let $x_1, \ldots, x_n \in \{0, 1\}$ and let $0 \le y \le n$ be an integer. Then Y has a binomial distribution with parameters n and θ. We may assume that $y = x_1 + \cdots + x_n$, otherwise there is nothing to show. Then

$$\begin{aligned}
&\mathbf{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \,|\, Y = y) = \frac{\mathbf{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n), Y = y)}{\mathbf{P}(Y = y)} \\
&= \frac{\mathbf{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n))}{\mathbf{P}(Y = y)} = \frac{\prod_{i=1}^n \mathbf{P}(X_i = x_i)}{\mathbf{P}(Y = y)} = \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} \\
&= \frac{\theta^y(1-\theta)^{n-y}}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} = \frac{1}{\binom{n}{y}} = \frac{1}{\binom{n}{x_1+\cdots+x_n}}.
\end{aligned}$$

Since the last expression does not depend on θ, Y is sufficient for θ.

Example 5.3. Let $X_1, \ldots, X_n$ be a random sample of size n from a Gaussian distribution with known variance $\sigma^2 > 0$ and unknown mean $\mu \in \mathbb{R}$. We claim that $Y := (X_1 + \cdots + X_n)/n$ is a sufficient statistic for μ. Let $x_1, \ldots, x_n \in \mathbb{R}$ and let $y \in \mathbb{R}$. Then Y is a Gaussian with variance σ^2/n and mean μ, and we may assume $y = (x_1 + \cdots + x_n)/n$, so that

$$\begin{aligned}
&f_{X_1,\ldots,X_n|Y}(x_1, \ldots, x_n|y) = \frac{f_{X_1,\ldots,X_n,Y}(x_1, \ldots, x_n, y)}{f_Y(y)} = \frac{f_{X_1,\ldots,X_n,Y}(x_1, \ldots, x_n, n^{-1}\sum_{i=1}^n x_i)}{f_Y(y)} \\
&= \frac{f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{f_Y(y)} = \frac{\sigma^{-n}(2\pi)^{-n/2}\exp\Big(-\frac{1}{2\sigma^2}(x_1^2 + \cdots + x_n^2) - \frac{n}{2\sigma^2}\mu^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\Big)}{n^{1/2}\sigma^{-1}(2\pi)^{-1/2}\exp\Big(-\frac{n}{2\sigma^2}y^2 - \frac{n}{2\sigma^2}\mu^2 + \frac{n\mu}{\sigma^2}y\Big)} \\
&= \frac{\sigma^{-n}(2\pi)^{-n/2}\exp\Big(-\frac{1}{2\sigma^2}(x_1^2 + \cdots + x_n^2)\Big)}{n^{1/2}\sigma^{-1}(2\pi)^{-1/2}\exp\Big(-\frac{n}{2\sigma^2}y^2\Big)}.
\end{aligned}$$

Since the last expression does not depend on μ, Y is sufficient for μ.

Theorem 5.4 (Factorization Theorem). *Suppose $X = (X_1, \dots, X_n)$ is a random sample of size n from a family $\{f_\theta \colon \theta \in \Theta\}$ of joint probability density functions, or a family of joint probability mass functions. (In the case of probability mass functions, we also assume that the set $\cup_{\theta\in\Theta}\{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$ is countable.) Let $t \colon \mathbb{R}^n \to \mathbb{R}^k$, so that $Y := t(X_1, \dots, X_n)$ is a statistic. Then Y is sufficient for θ if and only if there exist nonnegative functions $\{g_\theta \colon \theta \in \Theta\}$, $h \colon \mathbb{R}^n \to [0, \infty)$, $g_\theta \colon \mathbb{R}^k \to [0, \infty)$, such that*

$$f_\theta(x) = g_\theta(t(x))h(x), \qquad \forall\, \theta \in \Theta.$$

When $\{f_\theta \colon \theta \in \Theta\}$ are joint probability density functions, this equality holds for all $x \in \mathbb{R}^n$ except a set of measure zero. When $\{f_\theta \colon \theta \in \Theta\}$ are joint probability mass functions, this equality holds on the set $\cup_{\theta\in\Theta}\{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$.

A set $B \subseteq \mathbb{R}^n$ of measure zero satisfies: for all $\varepsilon > 0$, there exists a countable set of balls $B_1, B_2, \dots$ such that the total volume of $B_1, B_2, \dots$ is less than ε, and $B \subseteq \cup_{i=1}^\infty B_i$.

Proof. We only prove the case that the sampling distribution is discrete. The general case relies on measure theory, appearing in the Keener book, section 6.4.

Suppose Y is sufficient. Let $x \in \mathbb{R}^n$ and note that

$$f_\theta(x) = \mathbf{P}_\theta(X = x) = \mathbf{P}_\theta(X = x \text{ and } t(X) = t(x)) = \mathbf{P}_\theta(Y = t(x))\mathbf{P}_\theta(X = x | Y = t(x)).$$

By sufficiency, the last quantity does not depend on θ, so $f_\theta(x) = g_\theta(t(x))h(x)$, where $g_\theta(y) := \mathbf{P}_\theta(Y = y)$ for all $y \in \mathbb{R}^k$ and $h(x) := \mathbf{P}(X = x | Y = t(x))$ for all $x \in \mathbb{R}^n$.

Conversely, assume that $f_\theta(x) = g_\theta(t(x))h(x)$ as stated in the theorem. For any $x \in \mathbb{R}^n$, define $t^{-1}t(x) := \{y \in \mathbb{R}^n \colon t(y) = t(x)\}$. Then by our assumption and definitions

$$\begin{aligned}
\mathbf{P}_\theta(X = x | Y = t(x)) &= \frac{f_\theta(x)}{\mathbf{P}_\theta(Y = t(x))} = \frac{g_\theta(t(x))h(x)}{\mathbf{P}_\theta(t(X) = t(x))} = \frac{g_\theta(t(x))h(x)}{\sum_{z\in t^{-1}t(x)} \mathbf{P}_\theta(X = z)} \\
&= \frac{g_\theta(t(x))h(x)}{\sum_{z\in t^{-1}t(x)} f_\theta(z)} = \frac{g_\theta(t(x))h(x)}{\sum_{z\in t^{-1}t(x)} g_\theta(t(z))h(z)} \\
&= \frac{g_\theta(t(x))h(x)}{g_\theta(t(x)) \sum_{z\in t^{-1}t(x)} h(z)} = \frac{h(x)}{\sum_{z\in t^{-1}t(x)} h(z)}.
\end{aligned}$$

Since the probability does not depend on θ, Y is sufficient for θ. □

Remark 5.5. If $t(x) := x$ for all $x \in \mathbb{R}^n$, then the statistic $t(X_1, \dots, X_n) = (X_1, \dots, X_n)$ is automatically sufficient for θ, choosing $g_\theta := f_\theta$ and $h := 1$. So, at least one sufficient statistic always exists.

5.1.1. *Minimal Sufficient Statistics.* Suppose $t \colon \mathbb{R}^n \to \mathbb{R}^k$ and $Y := t(X_1, \dots, X_n)$ is a sufficient statistic for θ. Let $u \colon \mathbb{R}^n \to \mathbb{R}^m$ so that $Z := u(X_1, \dots, X_n)$ is another statistic. Suppose there exists $r \colon \mathbb{R}^m \to \mathbb{R}^k$ such that $r(u(x)) = t(x)$ for all $x \in \mathbb{R}^n$. That is, suppose

$$Y = r(Z).$$

It follows from the Factorization Theorem 5.4 that Z is also a sufficient statistic for θ since

$$f_\theta(x) = g_\theta(t(x))h(x) = g_\theta(r(u(x)))h(x), \qquad \forall\, x \in \mathbb{R}^n, \quad \forall\, \theta \in \Theta.$$

So, define $\widetilde{g}_\theta(y) := g_\theta(r(y))$ for all $y \in \mathbb{R}^m$. Then

$$f_\theta(x) = \widetilde{g}_\theta(u(x))h(x), \qquad \forall\, x \in \mathbb{R}^n, \quad \forall\, \theta \in \Theta.$$

That is, Z is sufficient for θ. That is, if Y is sufficient for θ, and if Y is a function of Z, then all "information" about θ is also stored in Z.

We would like to determine the parameter θ fitting the data using as little information as possible. For example, if we have a massive data set, we would like to use a minimal amount of memory on our computer in order to determine the parameter θ. The above observations motivate the following definition.

Definition 5.6 (Minimal Sufficient Statistic). Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a family $\{f_\theta \colon \theta \in \Theta\}$ of joint probability density functions, or joint probability mass functions. Let $t \colon \mathbb{R}^n \to \mathbb{R}^k$, so that $Y := t(X_1, \ldots, X_n)$ is a statistic. Assume that Y is sufficient for θ. Then Y is **minimal sufficient** for θ if, for every statistic $Z \colon \Omega \to \mathbb{R}^m$ that is sufficient for θ, there exists a function $r \colon \mathbb{R}^m \to \mathbb{R}^k$ such that $Y = r(Z)$.

Example 5.7. Let $X_1, \ldots, X_n$ be a random sample from a Gaussian distribution with unknown mean $\theta \in \mathbb{R}$ and variance 1. Then $\overline{X}$ is minimal sufficient for the mean θ. Let $t \colon \mathbb{R}^n \to \mathbb{R}^n$ so that $t(x) = x$ for all $x \in \mathbb{R}^n$ and define $Y := t(X_1, \ldots, X_n)$. Then Y is sufficient for θ, but Y is not minimal sufficient for θ.

It is fairly hard to prove directly that a statistic is minimal sufficient. The following theorem gives a condition for verifying minimal sufficiency that applies in particular to exponential families.

Theorem 5.8. *Suppose $(X_1, \ldots, X_n)$ is a random sample of size n from a family $\{f_\theta \colon \theta \in \Theta\}$ of joint probability density functions or joint probability mass functions. (In the case of probability mass functions, we also assume that the set $\cup_{\theta \in \Theta} \{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$ is countable.) Let $t \colon \mathbb{R}^n \to \mathbb{R}^m$ and define $Y := t(X_1, \ldots, X_n)$. When $\{f_\theta \colon \theta \in \Theta\}$ are joint probability density functions, suppose the following condition holds for every $x, y \in \mathbb{R}^n$, and when $\{f_\theta \colon \theta \in \Theta\}$ are joint probability mass functions, suppose the following condition holds for every x, y in the set $\cup_{\theta \in \Theta} \{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$.*

$$\begin{gathered} \exists\, c(x,y) \in \mathbb{R} \text{ that does not depend on } \theta \text{ such that} \\ f_\theta(x) = c(x,y) f_\theta(y) \quad \forall\, \theta \in \Theta \\ \text{if and only if } t(x) = t(y). \end{gathered}$$

Then Y is minimal sufficient.

Proof. We consider the PMF case only. We first prove sufficiency.

For any $z \in \{t(x) \colon x \in \mathbb{R}^n\}$, let x_z be any element of $t^{-1}z$, so that $t(x_z) = z$. Then for any $y \in \mathbb{R}^n$, $t(x_{t(y)}) = t(y)$, so by assumption, $f_\theta(y) = c(y, x_{t(y)}) f_\theta(x_{t(y)})$. So, define $g_\theta \colon \mathbb{R}^m \to \mathbb{R}$, $h \colon \mathbb{R}^n \to \mathbb{R}$ by

$$g_\theta(z) := f_\theta(x_z), \qquad h(y) := c(y, x_{t(y)}), \qquad \forall\, z \in \mathbb{R}^m,\ y \in \mathbb{R}^n.$$

Then $f_\theta(y) = g_\theta(t(y))h(y)$ for all $y \in \mathbb{R}^n$, so the Factorization Theorem 5.4 says that t is sufficient.

We now prove minimal sufficiency. Let $u \colon \mathbb{R}^n \to \mathbb{R}^m$ so that $Z := u(X_1, \ldots, X_n)$ is another statistic. Assume that Z is sufficient for θ. We are required to show that Y is a function of Z, i.e. t is a function of u. By the Factorization Theorem 5.4, $\exists\ h' \colon \mathbb{R}^m \to \mathbb{R}$ and $g'_\theta \colon \mathbb{R}^m \to \mathbb{R}$ for all $\theta \in \Theta$ such that, $\forall\ \theta \in \Theta$,

$$f_\theta(x) = g'_\theta(u(x)) h'(x), \qquad \forall\, x \in \mathbb{R}^n$$

Let $y \in \mathbb{R}^n$. If $h'(y) = 0$, then $f_\theta(y) = 0$ for all $\theta \in \Theta$. Our assumption ignores this case. That is, we may assume that $h'(y) \neq 0$. Let $x, y \in \mathbb{R}^n$ with $u(x) = u(y)$. Then

$$f_\theta(x) = g'_\theta(u(x))h'(x) = g'_\theta(u(y))h'(x) = g'_\theta(u(y))h'(y)\frac{h'(x)}{h'(y)} = f_\theta(y)\frac{h'(x)}{h'(y)}, \qquad \forall\, \theta \in \Theta.$$

So, define $c(x, y) := \frac{h'(x)}{h'(y)}$. We have

$$f_\theta(x) = f_\theta(y)c(x, y), \qquad \forall\, \theta \in \Theta.$$

So, by assumption $t(x) = t(y)$. That is, $u(x) = u(y)$ implies $t(x) = t(y)$. We conclude that t is a function of u by Exercise 5.9, so that Y is minimal sufficient for θ. □

Exercise 5.9. Let A, B, Ω be sets. Let $u\colon \Omega \to A$ and let $t\colon \Omega \to B$. Assume that, for every $x, y \in \Omega$, if $u(x) = u(y)$, then $t(x) = t(y)$. Show that there exists a function $s\colon A \to B$ such that

$$t = s(u).$$

Exercise 5.10. Let $\{f_\theta\colon \theta \in \Theta\}$ be a k-parameter exponential family $\{f_\theta\colon \theta \in \Theta,\, a(w(\theta)) < \infty\}$ of probability density functions or probability mass functions, where

$$f_\theta(x) := h(x) \exp\Big(\sum_{i=1}^{k} w_i(\theta)t_i(x) - a(w(\theta))\Big), \qquad \forall\, x \in \mathbb{R}.$$

For any $\theta \in \Theta$, let $w(\theta) := (w_1(\theta), \ldots, w_k(\theta))$. Assume that the following subset of $\mathbb{R}^k$ is k-dimensional:

$$\{w(\theta) - w(\theta') \in \mathbb{R}^k\colon \theta, \theta' \in \Theta\}.$$

That is, if $x \in \mathbb{R}^k$ satisfies $\langle x, y\rangle = 0$ for all y in this set, then $x = 0$. (Note that the assumption of the exercise is always satisfied for an exponential family in canonical form.)

Let $X = (X_1, \ldots, X_n)$ be a random sample of size n from f_θ. Define $t\colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$t(X) := \sum_{j=1}^{n} \left(t_1(X_j), \ldots, t_k(X_j)\right).$$

Show that $t(X)$ is minimal sufficient for θ. (Hint: if you get stuck, look at Example 3.12 in Keener.)

Conclude that if we sample from a Gaussian with unknown mean μ and variance $\sigma^2 > 0$, then $\overline{X}$ is minimal sufficient for θ and $(\overline{X}, S)$ is minimal sufficient for (μ, σ^2).

Warning: the f_θ exponential family mentioned here is a function of one variable. If you use the Theorem from class about checking the ratio of $f_\theta(x)/f_\theta(y)$, the functions there are *joint* density functions (i.e. the product of n copies of the same function).

Remark 5.11. If a minimal sufficient statistic exists, it is unique up to an invertible transformation. To see this, let $Y\colon \Omega \to \mathbb{R}^n$ and let $Z\colon \Omega \to \mathbb{R}^m$ be minimal sufficient statistics. By minimality of Y, there exists $r\colon \mathbb{R}^m \to \mathbb{R}^n$ such that $Y = r(Z)$. By minimality of Z, there exists $s\colon \mathbb{R}^n \to \mathbb{R}^m$ such that $Z = s(Y)$. Composing each of these identities with each other, we get

$$Y = r(s(Y)), \qquad Z = s(r(Z)).$$

That is, $r \circ s$ is the identity map on the range of Y, and $s \circ r$ is the identity map on the range of Z. That is, Y and Z are each the invertible image of each other.

The uniqueness of the minimal sufficient statistic is nice, since it implies that (up to an invertible map), there is at most one way to reduce the data at hand when we are trying to determine the parameter θ that fits our data.

Also, by this uniqueness, the converse of Theorem 5.8 should hold. The converse of Theorem 5.8 then suggests a strategy for proving existence of the minimal sufficient statistic that is used in the following Proposition.

Proposition 5.12 (**Existence of Minimal Sufficient Statistic**). *Suppose $X_1, \dots, X_n$ is a random sample of size n from a joint distribution f where $f \in \{f_\theta \colon \theta \in \Theta\}$ is a family of joint probability density functions, or a family of joint probability mass functions. (In the case of probability mass functions, we also assume that the set $\cup_{\theta\in\Theta}\{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$ is countable.) Then there exists a statistic Y that is minimal sufficient for θ.*

Proof. We first assume that Θ is countable. We relabel $\{f_\theta \colon \theta \in \Theta\}$ as $f_1, f_2, \dots$. Let $\mathbb{R}^{\mathbb{N}}/\sim$ denote $\mathbb{R}^{\mathbb{N}}$ where two elements $x, y \in \mathbb{R}^{\mathbb{N}}$ are considered equivalent if there exists $\alpha \in \mathbb{R}$ such that $x = \alpha y$. Define $t \colon \mathbb{R}^n \to \mathbb{R}^{\mathbb{N}}/\sim$ by

$$t(x) := \Big(f_1(x), f_2(x), \dots\Big).$$

We will show that $Y := t(X_1, \dots, X_n)$ is minimal sufficient for θ. This follows immediately from Theorem 5.8.

Proof of sufficiency for uncountable case. We take as given the following facts from real analysis. The set of functions $L := \{f \colon \mathbb{R} \to \mathbb{R} \colon \int_{\mathbb{R}} |f(x)|\,dx \le 1\}$ has a countable dense set, i.e. it has a countable subset L' such that for any $f \in L$, there exists a sequence $f_1, f_2, \dots \in L'$ such that $\lim_{i\to\infty} \int_{\mathbb{R}} |f_i(x) - f(x)|\,dx = 0$. Similarly, the set of functions $L := \{f \colon \mathbb{Z} \to \mathbb{R} \colon \sum_{x\in\mathbb{Z}} |f(x)| \le 1\}$ has a countable subset L' such that for any $f \in L$, there exists a sequence $f_1, f_2, \dots \in L'$ such that $\lim_{i\to\infty} \sum_{x\in\mathbb{Z}} |f_i(x) - f(x)| = 0$.

It follows that there is a countable set $\Theta' \subseteq \Theta$ such that, for any $\theta \in \Theta$ and for any $\varepsilon > 0$, there exists $\theta' \in \Theta'$ such that

$$\sup_{A\subseteq\Omega} |\mathbf{P}_\theta(A) - \mathbf{P}_{\theta'}(A)| < \varepsilon.$$

That is, $\{f_\theta \colon \theta' \in \Theta'\}$ is a countable dense subset of $\{f_\theta \colon \theta \in \Theta\}$. Without loss of generality, we label Θ' as $\{1, 2, 3, \dots\}$.

From above, we have a minimal sufficient statistic for $\theta \in \Theta'$. It remains to show this property extends to all $\theta \in \Theta$. To this end, we use regular conditional probabilities and the change of variables formula for the pushforward measure. $\forall\, \theta \in \Theta$, $\forall\, A \subseteq \mathbb{R}^n$, $B \subseteq \mathbb{R}^{\mathbb{N}}$,

$$\mathbf{P}_\theta(A \cap t^{-1}(B)) = \int_B \mathbf{P}_\theta(A|t = y)\mathbf{P}_\theta(t^{-1}(dy)) = \int_{t^{-1}(B)} \mathbf{P}_\theta(A|t = t(x))d\mathbf{P}_\theta(x).$$

Since Y is sufficient for $\theta \in \Theta'$, we can drop the θ subscript on the right to get

$$\mathbf{P}_i(A \cap t^{-1}(B)) = \int_{t^{-1}(B)} \mathbf{P}(A|t = t(x))d\mathbf{P}_i(x), \qquad \forall\, i \ge 1,\ A \subseteq \mathbb{R}^n,\ B \subseteq \mathbb{R}^{\mathbb{N}}.$$

Or, written in analytic form,

$$\int_{A\cap t^{-1}(B)} d\mathbf{P}_i(x) = \int_{t^{-1}(B)} \mathbf{P}(A|t = t(x))d\mathbf{P}_i(x), \qquad \forall\, i \ge 1,\ A \subseteq \mathbb{R}^n,\ B \subseteq \mathbb{R}^{\mathbb{N}}. \qquad (*)$$

This condition is preserved under total variation limits, so it extends to the whole set of densities. That is, for any $\theta \in \Theta$, we choose $\theta_1, \theta_2, \ldots \in \Theta'$ such that

$$\lim_{i\to\infty} \sup_{A\subseteq\mathbb{R}^n} |\mathbf{P}_{\theta_i}(A) - \mathbf{P}_\theta(A)| = 0.$$

Then

$$\left|\int_{t^{-1}B} \mathbf{P}(A|t=t(x))d\mathbf{P}_i(x) - \int_{t^{-1}B} \mathbf{P}_\theta(A|t=t(x))d\mathbf{P}_\theta(x)\right| \le \left|\mathbf{P}_i(t^{-1}B) - \mathbf{P}_\theta(t^{-1}B)\right| \to 0,$$

as $i \to \infty$. Similarly, as $i \to \infty$

$$\left|\int_{A\cap t^{-1}(B)} d\mathbf{P}_i(x) - \int_{A\cap t^{-1}(B)} d\mathbf{P}_\theta(x)\right| \to 0.$$

We conclude by $(*)$ that, for all $\theta \in \Theta$,

$$\int_{A\cap t^{-1}(B)} d\mathbf{P}_\theta(x) = \int_{t^{-1}(B)} \mathbf{P}(A|t=t(x))d\mathbf{P}_\theta(x), \qquad \forall\, A \subseteq \mathbb{R}^n,\ B \subseteq \mathbb{R}^{\mathbb{N}}.$$

So, going backwards and using the regular conditional probability definition of $\mathbf{P}_\theta$, we have

$$\int_{t^{-1}(B)} \mathbf{P}_\theta(A|t=t(x))d\mathbf{P}_\theta(x) = \int_{t^{-1}(B)} \mathbf{P}(A|t=t(x))d\mathbf{P}_\theta(x), \qquad \forall\, A \subseteq \mathbb{R}^n,\ B \subseteq \mathbb{R}^{\mathbb{N}}.$$

That is, $\mathbf{P}_\theta(A|t=y)$ does not depend on $\theta \in \Theta$. Therefore, Y is sufficient for $\theta \in \Theta$.

Proof of minimal sufficiency for uncountable case. Minimal sufficiency follows by the proof of Theorem 5.8.

□

Exercise 5.13. Let $\mathbf{P}_1, \mathbf{P}_2$ be two probability laws on the sample space $\Omega = \mathbb{R}$. Suppose these laws have densities $f_1, f_2 \colon \mathbb{R} \to [0,\infty)$ so that

$$\mathbf{P}_i(A) = \int_A f_i(x)dx, \qquad \forall\, i = 1,2, \qquad \forall\, A \subseteq \mathbb{R}.$$

Show that

$$\sup_{A\subseteq\mathbb{R}} |\mathbf{P}_1(A) - \mathbf{P}_2(A)| = \frac{1}{2}\int_{\mathbb{R}} |f_1(x) - f_2(x)|\, dx.$$

(Hint: consider $A := \{x \in \mathbb{R} \colon f_1(x) > f_2(x)\}$.)

Similarly, if $\mathbf{P}_1, \mathbf{P}_2$ are probability laws on $\Omega = \mathbb{Z}$, show that

$$\sup_{A\subseteq\mathbb{Z}} |\mathbf{P}_1(A) - \mathbf{P}_2(A)| = \frac{1}{2}\sum_{z\in\mathbb{Z}} |\mathbf{P}_1(z) - \mathbf{P}_2(z)|\,.$$

5.2. **Ancillary Statistics.** Minimal sufficient statistics provide sufficient information to estimate a parameter θ in a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. However, even a minimal sufficient statistic can have excess "information." For example, we saw in Proposition 5.12 that a minimal sufficient statistic can have infinitely many nontrivial components in its range. It would be desirable to come up with statistics that contain as little unnecessary information as possible, while still being minimal sufficient. In order to accomplish this task, we first define what we mean by "excess information" of a statistic.

Definition 5.14 (**Ancillary Statistic**). Suppose $X_1, \ldots, X_n$ is a random sample of size n from a distribution f where $f \in \{f_\theta \colon \theta \in \Theta\}$ is a family of distributions. A statistic $Y = t(X_1, \ldots, X_n)$, $t \colon \mathbb{R}^n \to \mathbb{R}^m$ is **ancillary** for θ if the distribution of Y does not depend on θ.

Example 5.15. Let $X_1, \ldots, X_n$ be a random sample of size n from the location family for the Cauchy distribution:

$$f_\theta(x) := \prod_{i=1}^n \frac{1}{\pi} \frac{1}{1 + (x_i - \theta)^2}, \qquad \forall\, x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad \forall\, \theta \in \mathbb{R}.$$

Then the order statistics $X_{(1)} \leq \cdots \leq X_{(n)}$ are minimal sufficient for θ. Sufficiency follows by the Factorization Theorem 5.4 since, if $t(x) := (x_{(1)}, \ldots, x_{(n)})$, then $f_\theta(t(x)) = f_\theta(x)$. Minimal sufficiency follows from Theorem 5.8, since if $x, y \in \mathbb{R}^n$ are fixed, then the following ratio is constant in θ

$$\frac{f_\theta(x)}{f_\theta(y)} = \frac{\prod_{i=1}^n \frac{1}{1+(x_i-\theta)^2}}{\prod_{i=1}^n \frac{1}{1+(y_i-\theta)^2}} = \frac{\prod_{i=1}^n [1 + (y_i - \theta)^2]}{\prod_{i=1}^n [1 + (x_i - \theta)^2]},$$

only when $t(x) = t(y)$. To see this, note that the top and bottom are each polynomials in θ, and these polynomials must be a constant multiple of each other, so their (complex) roots must be identical (counting multiplicities), and these roots are $\theta = x_i \pm \sqrt{-1}$, $\theta = y_i \pm \sqrt{-1}$ respectively ($i = 1, \ldots, n$), so that $t(x) = t(y)$.

Even though the order statistics $(X_{(1)}, \ldots, X_{(n)})$ are minimal sufficient for θ in this case, they certainly seem to contain a lot of extraneous information about θ. Indeed, the statistic $X_{(n)} - X_{(1)}$ is ancillary. To see this, let $Z_1, \ldots, Z_n$ be independent Cauchy random variables, i.e. they each have density $\frac{1}{\pi} \frac{1}{1+x^2}$ for all $x \in \mathbb{R}$. Then $X_i = Z_i + \theta$ for all $1 \leq i \leq n$, so that $X_{(i)} = Z_{(i)} + \theta$ for all $1 \leq i \leq n$, so that $X_{(n)} - X_{(1)} = Z_{(n)} - Z_{(1)}$, and the last expression does not depend on θ.

5.3. **Complete Statistics.** Continuing Example 5.15, we saw that $X_{(n)} - X_{(1)}$ is ancillary, i.e. there exists a constant c that does not depend on θ such that

$$\mathbf{E}_\theta[[X_{(n)} - X_{(1)}] 1_{\{0 \leq X_{(n)} - X_{(1)} \leq 1\}} - c] = 0, \qquad \forall\, \theta \in \Theta = \mathbb{R}.$$

Let $f \colon \mathbb{R}^n \to \mathbb{R}$ so that $f(x_1, \ldots, x_n) := [x_n - x_1] 1_{0 \leq x_n - x_1 \leq 1} - c$ for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Then we have shown that

$$\mathbf{E}_\theta f(Y) = 0, \qquad \forall\, \theta \in \Theta,$$

where $Y = (X_{(1)}, \ldots, X_{(n)})$ is the vector of order statistics. Note that

$$f(Y) = [X_{(n)} - X_{(1)}] 1_{\{0 \leq X_{(n)} - X_{(1)} \leq 1\}} - c \neq 0.$$

The above observations imply that Y contains extraneous information, despite it being minimal sufficient. That is, Y is **not** complete, in the following sense.

Definition 5.16 (**Complete Statistic**). Suppose $X_1, \ldots, X_n$ is a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. Let $t \colon \mathbb{R}^n \to \mathbb{R}^m$. A statistic $Y = t(X_1, \ldots, X_n)$ is **complete** for $\{f_\theta \colon \theta \in \Theta\}$ if the following holds:

For any $f \colon \mathbb{R}^m \to \mathbb{R}$ such that $\mathbf{E}_\theta f(Y) = 0 \quad \forall\, \theta \in \Theta$, it holds that $f(Y) = 0$.

(When we assume that $\mathbf{E}_\theta f(Y)$ can be defined, we are also assuming, as usual, that $\mathbf{E}_\theta |f(Y)| < \infty$ for all $\theta \in \Theta$.)

Remark 5.17. From our discussion above, we see that a nonconstant complete statistic is not ancillary. (If Y is ancillary, then there is a constant $c \in \mathbb{R}$ such that $\mathbf{E}_\theta(Y - c) = 0$ for all $\theta \in \Theta$, and if Y is also complete, we then have $Y - c = 0$, so that $Y = c$.) Also, a complete statistic may not be sufficient. Consider for example a statistic that is constant.

Remark 5.18. Unfortunately, a complete sufficient statistic might not exist.

Exercise 5.19. Give an example of a statistic Y that is complete and nonconstant, but such that Y is not sufficient.

Exercise 5.20. This exercise shows that a complete sufficient statistic might not exist.

Let $X_1, \dots, X_n$ be a random sample of size n from the uniform distribution on the three points $\{\theta, \theta + 1, \theta + 2\}$, where $\theta \in \mathbb{R}$.

- Show that the vector $Y := (X_{(1)}, X_{(n)})$ is minimal sufficient for θ.
- Show that Y is not complete by considering $X_{(n)} - X_{(1)}$.
- Using minimal sufficiency, conclude that any sufficient statistic for θ is not complete.

Example 5.21. We return to Example 5.2. Let $X = (X_1, \dots, X_n)$ be a random sample of size n from a Bernoulli distribution with parameter $0 < \theta < 1$. We showed that $Y := X_1 + \cdots + X_n$ is a sufficient statistic for θ. We show now that Y is also complete. Let $f \colon \mathbb{R} \to \mathbb{R}$. Assume that $\mathbf{E}_\theta f(Y) = 0$. Since Y is binomial, we can write out this expected value as the following sum

$$0 = \mathbf{E}_\theta f(Y) = \sum_{j=0}^{n} f(j) \binom{n}{j} \theta^j (1-\theta)^{n-j}, \qquad \forall\, \theta \in (0,1).$$

Let $\alpha := \theta/(1-\theta)$. Dividing by θ^n and rewriting this expression, we have

$$0 = \sum_{j=0}^{n} f(j) \binom{n}{j} \alpha^j, \qquad \forall\, \alpha > 0.$$

This function of α is a polynomial. A polynomial can only be zero for all $\alpha > 0$ if the polynomial itself is always zero. That is, $f(j) = 0$ for all $0 \le j \le n$. Therefore, Y is complete.

Example 5.22. We return to Example 5.3. Let $X_1, \dots, X_n$ be a random sample of size n from a Gaussian distribution with known variance $\sigma^2 > 0$ and unknown mean $\mu \in \mathbb{R}$. We claim that $Y := (X_1 + \cdots + X_n)/n$ is a complete sufficient statistic for μ. For simplicity of notation we just consider $n = \sigma = 1$, so that $Y = X_1$ is a Gaussian random variable. Let $f \colon \mathbb{R} \to \mathbb{R}$ such that $\mathbf{E}_\mu |f(Y)| < \infty$ for all $\mu \in \mathbb{R}$ and such that

$$0 = \mathbf{E}_\mu f(Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(y) e^{-\frac{1}{2}(y-\mu)^2} dy, \qquad \forall\, \mu \in \mathbb{R}.$$

Multiplying by $e^{\mu^2/2}\sqrt{2\pi}$, we equivalently have

$$0 = \int_{\mathbb{R}} f(y) e^{-\frac{1}{2}y^2} e^{y\mu} dy, \qquad \forall\, \mu \in \mathbb{R}.$$

In case $f(y) \ge 0$ for all $y \in \mathbb{R}$, we must have $f = 0$ since the integral of a nonnegative function being zero implies that the function is zero. In case f is positive somewhere and

negative elsewhere, write $f = f_+ - f_-$ where $f_+ := \max(f, 0)$ and $f_- = \max(-f, 0)$. Then, using the above equality for any μ and dividing by the case when $\mu = 0$, we get

$$\frac{\int_{\mathbb{R}} f_+(y) e^{-\frac{1}{2}y^2} e^{y\mu} dy}{\int_{\mathbb{R}} f_+(y) e^{-\frac{1}{2}y^2} dy} = \frac{\int_{\mathbb{R}} f_-(y) e^{-\frac{1}{2}y^2} e^{y\mu} dy}{\int_{\mathbb{R}} f_-(y) e^{-\frac{1}{2}y^2} dy}, \qquad \forall\, \mu \in \mathbb{R}.$$

The left side is the moment generating function of a random variable with density $\frac{f_+(x)e^{-\frac{1}{2}x^2}}{\int_{\mathbb{R}} f_+(y)e^{-\frac{1}{2}y^2}dy}$, and similarly for the right side with f_-. Inverting the moment generating function by Theorem 11.2, we conclude that $f_+ = f_-$, therefore $f = 0$ as desired.

Exercise 5.23 ((Optional) This exercise requires some measure theory so it is optional.)**.** Let $\{f_\theta \colon \theta \in \Theta\}$ be a k-parameter exponential family $\{f_\theta \colon \theta \in \Theta,\, a(w(\theta)) < \infty\}$ of **joint** probability density functions or probability mass functions in canonical form, where

$$f_w(x) := h(x) \exp\Big(\sum_{i=1}^{k} w_i t_i(x) - a(w) \Big), \qquad \forall\, x \in \mathbb{R}^n, \qquad \forall\, w \in \{w \in \mathbb{R}^k \colon a(w) < \infty\}.$$

Assume that the following subset of $\mathbb{R}^k$ contains an open set in $\mathbb{R}^k$:

$$\{w \in \mathbb{R}^k \colon a(w) < \infty\}.$$

Assume also that there is no redundancy in the functions $t_1, \ldots, t_k$, i.e. assume: if $\exists$ $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ such that $\sum_{i=1}^k \alpha_i t_i(x) = 0$ for all $x \in \mathbb{R}^n$, then $\alpha_1 = \cdots = \alpha_k = 0$.

Let X be a random sample **of size** 1 from f_θ (so $X = (X_1, \ldots, X_n)$, and $X_1, \ldots, X_n$ are all real valued). Define $t \colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$t(X) := (t_1(X), \ldots, t_k(X)).$$

Show that $t(X)$ is complete for θ.

Hint: if you get stuck, look at Theorem 4.3.1 in Lehmann-Romano. An early step in the proof uses the change of variables formula for the pushforward measure.

Once we know the above statement, we can deduce the following about repeated random samples from a single variable exponential family.

Let $\{f_\theta \colon \theta \in \Theta\}$ be a k-parameter exponential family $\{f_\theta \colon \theta \in \Theta,\, a(w(\theta)) < \infty\}$ of probability density functions or probability mass functions in canonical form, where

$$f_w(x) := h(x) \exp\Big(\sum_{i=1}^{k} w_i t_i(x) - a(w) \Big), \qquad \forall\, x \in \mathbb{R}, \qquad \forall\, w \in \{w \in \mathbb{R}^k \colon a(w) < \infty\}.$$

Assume that the following subset of $\mathbb{R}^k$ contains an open set in $\mathbb{R}^k$:

$$\{w \in \mathbb{R}^k \colon a(w) < \infty\}.$$

Assume also that there is no redundancy in the functions $t_1, \ldots, t_k$, i.e. assume: if $\exists$ $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ such that $\sum_{i=1}^k \alpha_i t_i(x) = 0$ for all $x \in \mathbb{R}$, then $\alpha_1 = \cdots = \alpha_k = 0$.

Let $X_1, \ldots, X_n$ be a random sample **of size** n from f_θ. Define $t \colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$t(X) := \sum_{j=1}^{n} (t_1(X_j), \ldots, t_k(X_j)).$$

Show that $t(X)$ is complete for θ.

Exercise 5.24 (**Conditional Expectation as a Random Variable**). Let $X, Y, Z\colon \Omega \to \mathbb{R}$ be discrete or continuous random variables. Let A be the range of Y. Define $g\colon A \to \mathbb{R}$ by $g(y) := \mathbf{E}(X|Y = y)$, for any $y \in A$. We then define the **conditional expectation** of X given Y, denoted $\mathbf{E}(X|Y)$, to be the random variable $g(Y)$.

(i) Let X, Y be random variables such that (X, Y) is uniformly distributed on the triangle $\{(x, y) \in \mathbb{R}^2\colon x \geq 0, y \geq 0, x + y \leq 1\}$. Show that

$$\mathbf{E}(X|Y) = \frac{1}{2}(1 - Y).$$

(ii) Prove the following version of the Total Expectation Theorem

$$\mathbf{E}(\mathbf{E}(X|Y)) = \mathbf{E}(X).$$

- If X is a random variable, and if $f(t) := \mathbf{E}(X - t)^2$, $t \in \mathbb{R}$, then the function $f\colon \mathbb{R} \to \mathbb{R}$ is uniquely minimized when $t = \mathbf{E}X$. A similar minimizing property holds for conditional expectation. Let $h\colon \mathbb{R} \to \mathbb{R}$. Show that the quantity $\mathbf{E}(X - h(Y))^2$ is minimized among all functions $h\colon \mathbb{R} \to \mathbb{R}$ when $h(Y) = \mathbf{E}(X|Y)$. (Hint: use the previous item.)

(iii) Show the following:

$$\mathbf{E}(Xh(Y)|Y) = h(Y)\mathbf{E}(X|Y).$$
$$\mathbf{E}([\mathbf{E}(X|h(Y))]\,|Y) = \mathbf{E}(X|h(Y)).$$

(iv) Show the following

$$\mathbf{E}(X|X) = X.$$
$$\mathbf{E}(X + Y|Z) = \mathbf{E}(X|Z) + \mathbf{E}(Y|Z).$$

(v) If Z is independent of X and Y, show that

$$\mathbf{E}(X|Y, Z) = \mathbf{E}(X|Y).$$

(Here $\mathbf{E}(X|Y, Z)$ is notation for $\mathbf{E}(X|(Y, Z))$ where (Y, Z) is interpreted as a random vector, so that X is conditioned on the random vector (Y, Z).)

Exercise 5.23 strengthens Exercise 5.10 by the following Theorem.

Theorem 5.25 (**Bahadur's Theorem**). *If Y is a complete sufficient statistic for a family $\{f_\theta\colon \theta \in \Theta\}$ of joint probability densities or joint probability mass functions, then Y is a minimal sufficient statistic for θ. (In the case of probability mass functions, we also assume that the set $\cup_{\theta\in\Theta}\{x \in \mathbb{R}^n\colon f_\theta(x) > 0\}$ is countable.)*

Remark 5.26. So, by Remark 5.11, a complete sufficient statistic is unique, up to an invertible map. Also, by Example 5.15, the converse of Bahadur's Theorem is false.

Proof. By Proposition 5.12, there exists a minimal sufficient statistic Z for θ. To show that Y is minimal sufficient, it suffices to find a function r such that $Y = r(Z)$. Define $r(Z) := \mathbf{E}_\theta(Y|Z)$. Since Z is minimal sufficient and Y is sufficient by assumption, there exists a function u such that $Z = u(Y)$. So, using the definition of $r(Z)$, we have by Exercise 5.24

$$\mathbf{E}_\theta(r(u(Y))) = \mathbf{E}_\theta r(Z) = \mathbf{E}_\theta\mathbf{E}_\theta(Y|Z) \overset{(ii)}{=} \mathbf{E}_\theta Y.$$

That is,

$$\mathbf{E}_\theta[r(u(Y)) - Y] = 0, \qquad \forall\, \theta \in \Theta.$$

Since Y is complete, we conclude that $r(u(Y)) = Y$, and since $r(u(Y)) = r(Z)$, we have $r(Z) = Y$, as desired. □

The following theorem says that complete sufficient statistics have no ancillary information, unlike the minimal sufficient statistics, as we saw in Example 5.15.

Theorem 5.27 (**Basu's Theorem**). *If Y is a complete sufficient statistic for $\{f_\theta \colon \theta \in \Theta\}$, and if Z is ancillary for θ, then for all $\theta \in \Theta$, Y and Z are independent with respect to f_θ.*

Proof. Suppose $Y \colon \Omega \to \mathbb{R}^k$ and $Z \colon \Omega \to \mathbb{R}^m$. Let $A \subseteq \mathbb{R}^k$ and let $B \subseteq \mathbb{R}^m$. We need to show that

$$\mathbf{P}_\theta(Y \in A, Z \in B) = \mathbf{P}_\theta(Y \in A)\mathbf{P}_\theta(Z \in B), \qquad \forall\, \theta \in \Theta.$$

Using Exercise 5.24,

$$\mathbf{P}_\theta(Y \in A, Z \in B) = \mathbf{E}_\theta 1_{Y\in A} 1_{Z\in B} \overset{(ii)}{=} \mathbf{E}_\theta \mathbf{E}_\theta(1_{Y\in A}1_{Z\in B}|Y) \overset{(iii)}{=} \mathbf{E}_\theta 1_{Y\in A}\mathbf{E}_\theta(1_{Z\in B}|Y)$$

Let $g(Y) := \mathbf{E}_\theta(1_{Z\in B}|Y)$. Note that $g(Y)$ does not depend on θ by Exercise 6.9, i.e. $g(Y)$ is a function of the sample but not an explicit function of θ. By Exercise 5.24 (ii),

$$\mathbf{E}_\theta g(Y) = \mathbf{E}_\theta \mathbf{E}_\theta(1_{Z\in B}|Y) = \mathbf{E}_\theta 1_{Z\in B} = \mathbf{P}_\theta(Z \in B).$$

The quantity $c := \mathbf{P}_\theta(Z \in B)$ does not depend on θ since Z is ancillary. Then $\mathbf{E}_\theta[g(Y) - c] = 0$ for all $\theta \in \Theta$. Since Y is complete, we must have $g(Y) = c$. In conclusion,

$$\mathbf{P}_\theta(Y \in A, Z \in B) = \mathbf{E}_\theta 1_{Y\in A}\mathbf{P}_\theta(Z \in B) = \mathbf{P}_\theta(Y \in Z)\mathbf{P}_\theta(Z \in B).$$

□

So, Basu's Theorem says that complete sufficient statistics provide an "optimal" reduction of data. Unfortunately, as we saw above, complete sufficient statistics might not exist, so we might not be able to reduce data in this way.

5.4. **Additional Comments.** Sufficient and ancillary statistics were introduced by Fisher in 1920. Complete and minimal sufficient statistics were studied in the mid 1900s by Bahadur, Halmos, and Savage, and Lehmann and Scheffé.

Above, we have typically focused on families of probability density functions or probability mass functions, in order to avoid use of measure theory. However, many of the above theorems naturally generalize to the setting of a dominated family of functions.

Definition 5.28 (**Dominated Family**). Let $\Theta \subseteq \mathbb{R}^m$. Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of functions so that $f_\theta \colon \mathbb{R}^n \to [0, \infty)$ for all $\theta \in \Theta$. We say that $\{f_\theta \colon \theta \in \Theta\}$ is a **dominated family** if there exists a measure μ on $\mathbb{R}^m$ such that $\mathbf{P}_\theta$ is absolutely continuous with respect to μ, for all $\theta \in \Theta$.

For example, a family of probability density functions is absolutely continuous with respect to Lebesgue measure. And a family of probability mass functions is absolutely continuous with respect to a counting measure, if $\cup_{\theta\in\Theta}\{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$ is countable.

We can then restate the Factorization Theorem and its Corollaries for dominated families.

Theorem 5.29 (**Factorization Theorem**). *Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a dominated family $\{f_\theta \colon \theta \in \Theta\}$ that is dominated by a measure μ on $\mathbb{R}^n$. That is, $f_\theta \colon \mathbb{R}^n \to [0, \infty)$ for all $\theta \in \Theta$. Let $t \colon \mathbb{R}^n \to \mathbb{R}^k$, so that $Y := t(X_1, \ldots, X_n)$*

is a statistic. Then Y is sufficient for θ if and only if there exist nonnegative functions $\{g_\theta \colon \theta \in \Theta\}$, $h \colon \mathbb{R}^n \to [0,\infty)$, $g_\theta \colon \mathbb{R}^k \to [0,\infty)$, such that

$$f_\theta(x) = g_\theta(t(x))h(x), \qquad \forall\, \theta \in \Theta, \quad \text{for a.e. } x \text{ with respect to } \mu.$$

Theorem 5.30. *Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a dominated family $\{f_\theta \colon \theta \in \Theta\}$ that is dominated by a measure μ on $\mathbb{R}^n$. Let $t \colon \mathbb{R}^n \to \mathbb{R}^m$ and define $Y := t(X_1, \ldots, X_n)$. Suppose the following condition holds for a.e. $x, y \in \mathbb{R}^n$ with respect to μ:*

$$\exists\, c(x,y) \in \mathbb{R} \text{ that does not depend on } \theta \text{ such that}$$
$$f_\theta(x) = c(x,y) f_\theta(y) \quad \forall\, \theta \in \Theta$$
$$\text{if and only if } t(x) = t(y).$$

Then Y is minimal sufficient.

Proposition 5.31 (**Existence of Minimal Sufficient Statistic**)**.** *Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a dominated family $\{f_\theta \colon \theta \in \Theta\}$ that is dominated by a measure μ on $\mathbb{R}^n$. Suppose the set $\{f_\theta \colon \theta \in \Theta\}$ has a countable dense set with respect to the total variation metric $d(f_\theta, f_{\theta'}) = \sup_{B \subseteq \mathbb{R}^n} |\mathbf{P}_\theta(B) - \mathbf{P}_{\theta'}(B)|$. Then there exists a statistic Y that is minimal sufficient for θ.*

To see the original proof, read Theorem 6.1 in "Completeness, Similar Regions, and Unbiased Estimation-Part I" by Lehmann and Scheffé.

6. Point Estimation

Let $X_1, \ldots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. If Y is a statistic that is used to estimate the parameter θ that fits the data at hand, we then refer to Y as a **point estimator** or **estimator**

6.1. Heuristic Principles for Finding Good Estimators.

Let $X_1, \ldots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$.

Definition 6.1 (**Likelihood**)**.** If we have data $x \in \mathbb{R}^n$, then the function $\ell \colon \Theta \to [0,\infty)$ defined by

$$\ell(\theta) := f_\theta(x)$$

is called the **likelihood function**

Likelihood Principle. All data relevant to estimating the parameter θ is contained in the likelihood function. (This is the intuition in the proof of Proposition 5.12.)

Sufficiency Principle. If $Y = t(X_1, \ldots, X_n)$ is a sufficient statistic, and if we have two results $x, y \in \mathbb{R}^n$ from an experiment with the same statistics $t(x) = t(y)$, then our estimate of the parameter θ should be the same for either experimental result. (This is the intuition behind Theorem 5.8.)

Equivariance Principle. If the family of distributions $\{f_\theta \colon \theta \in \Theta\}$ is invariant under some symmetry, then the estimator of θ should respect the same symmetry.

For example, a location family is invariant under translation, so an estimator for the location parameter should commute with translations.

6.2. **Evaluating Estimators.** There are many different ways to create estimators. A priori, it might not be clear which estimator is the best. One desirable property of an estimator is that it is unbiased.

Definition 6.2. Let $X_1, \dots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. Let $t \colon \mathbb{R}^n \to \mathbb{R}^k$ and let $Y := t(X_1, \dots, X_n)$ be an estimator for $g(\theta)$. Let $g \colon \Theta \to \mathbb{R}^k$. We say that Y is **unbiased** for $g(\theta)$ if

$$\mathbf{E}_\theta Y = g(\theta), \qquad \forall\, \theta \in \Theta.$$

For example, we saw in Exercise 4.5 that the sample mean and sample variance are unbiased estimates of the mean and variance, respectively.

Even if an estimator is unbiased, its distribution of values might be quite far from θ. Recall that we made a similar observation that the Law of Large Numbers does not give any information about the Central Limit Theorem. It is desirable to examine the distribution of values of the estimator. The most common way to check the quality of an estimator in this sense is to examine the mean-squared error, or squared L_2 norm, of the estimator minus θ:

$$\mathbf{E}_\theta (Y - g(\theta))^2.$$

If the estimator is unbiased, this quantity is equal to the variance of Y.

Definition 6.3 (**UMVU**)**.** Let $X_1, \dots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. Let $g \colon \Theta \to \mathbb{R}$. Let $t \colon \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X_1, \dots, X_n)$ be an unbiased estimator for $g(\theta)$. We say that Y is **uniformly minimum variance unbiased (UMVU)** if, for any other unbiased estimator Z for $g(\theta)$, we have

$$\mathrm{Var}_\theta(Y) \le \mathrm{Var}_\theta(Z), \qquad \forall\, \theta \in \Theta.$$

More generally, we are given a **loss function**

$$\ell \colon \Theta \times \mathbb{R}^k \to \mathbb{R},$$

and we could be asked to minimize the **risk function**

$$r(\theta, Y) := \mathbf{E}_\theta \ell(\theta, Y)$$

over all possible estimators Y. In the case of mean-squared error, we have $\ell(\theta, y) := (y - g(\theta))^2$ for all $y, \theta \in \mathbb{R}$.

Definition 6.4 (**UMRU**)**.** Let $X_1, \dots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. Let $g \colon \Theta \to \mathbb{R}$. Let $t \colon \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X_1, \dots, X_n)$ be an unbiased estimator for $g(\theta)$. We say that Y is **uniformly minimum risk unbiased (UMRU)** for risk function r if, for any other unbiased estimator Z for $g(\theta)$, we have

$$r(\theta, Y) \le r(\theta, Z), \qquad \forall\, \theta \in \Theta.$$

Remark 6.5. Unfortunately the UMVU might not exist. Suppose we want a UMVU for a binomial random variable X with known parameter n and unknown parameter $0 < \theta < 1$, and we want an estimator for $\theta/(1-\theta)$. In fact, no unbiased estimate exists for this function, since $\mathbf{E}_\theta t(X) = \sum_{j=0}^n \binom{n}{j} t(j) \theta^j (1-\theta)^{n-j}$ and this is a polynomial in θ, i.e. only polynomials in θ of degree at most n can possible have unbiased estimates. And $\theta/(1-\theta)$ is not a polynomial in θ.

Recall that the function $t \mapsto t^2$ is a convex function of t.

Definition 6.6. Let $\phi\colon \mathbb{R} \to \mathbb{R}$. We say that ϕ is **strictly convex** if, for any $x, y \in \mathbb{R}$ with $x \neq y$ and for any $t \in (0,1)$, we have

$$\phi(tx + (1-t)y) < t\phi(x) + (1-t)\phi(y).$$

A strictly convex function is convex.

The Rao-Blackwell Theorem says that, if $\ell(\theta, y)$ is convex in y, then any sufficient statistic can be used to improve any estimator for $g(\theta)$.

Theorem 6.7 (**Rao-Blackwell**). *Let Z be a sufficient statistic for $\{f_\theta\colon \theta \in \Theta\}$ and let Y be an estimator for $g(\theta)$. Define $W := \mathbf{E}_\theta(Y|Z)$. (Since Z is sufficient for θ, W does not depend on θ by Exercise 6.9, i.e. W is a well-defined function of the random sample but not an explicit function of θ.) Let $\theta \in \Theta$ with $r(\theta, Y) < \infty$ and such that $\ell(\theta, y)$ is convex in $y \in \mathbb{R}$. Then*

$$r(\theta, W) \leq r(\theta, Y).$$

And if $\ell(\theta, y)$ is strictly convex in y, then this inequality is strict unless $W = Y$.

Proof. By the (conditional) Jensen's inequality, Exercise 6.8

$$\ell(\theta, W) = \ell(\theta, \mathbf{E}_\theta(Y|Z)) \leq \mathbf{E}_\theta[\ell(\theta, Y)|Z].$$

Taking expected values of both sides and applying Exercise 5.24(ii), we get

$$r(\theta, W) \leq \mathbf{E}_\theta\mathbf{E}_\theta[\ell(\theta, Y)|Z] = \mathbf{E}_\theta\ell(\theta, Y) = r(\theta, Y).$$

And if $\ell(\theta, y)$ is strictly convex in y, then this inequality is strict, unless Y is a function of Z. If Y is a function of Z, then $\mathbf{E}_\theta(Y|Z) = Y$, so $W = Y$. □

From Exercise 5.24(iii), if Y is a function of Z, then $\mathbf{E}_\theta(Y|Z) = Y$. Conversely, if $\mathbf{E}_\theta(Y|Z) = Y$, then by definition of $\mathbf{E}_\theta(Y|Z)$, we conclude that Y is a function of Z. That is, $\mathbf{E}_\theta(Y|Z) = Y$ if and only if Y is a function of Z.

Exercise 6.8 (**Conditional Jensen Inequality**)**.** Prove Jensen's inequality for the conditional expectation. Let $X, Y\colon \Omega \to \mathbb{R}$ be random variables that are either both discrete or both continuous. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be convex. Then

$$\phi(\mathbf{E}(X|Y)) \leq \mathbf{E}(\phi(X)|Y)$$

If ϕ is strictly convex, then equality holds only if X is constant on any set where Y is constant. That is, (by Exercise 5.9) equality holds only if X is a function of Y.

(Hint: first show that if $X \geq Z$ then $\mathbf{E}(X|Y) \geq \mathbf{E}(Z|Y)$.)

Exercise 6.9. Let Y, Z be a statistics, and suppose Z is sufficient for $\{f_\theta\colon \theta \in \Theta\}$. Show that $W := \mathbf{E}_\theta(Y|Z)$ does not depend on θ. That is, there is a function $t\colon \mathbb{R}^n \to \mathbb{R}$ that does not depend on θ such that $W = t(X)$, where X is the sample distribution.

Remark 6.10. By Exercise 5.24, if Y is unbiased, then $\mathbf{E}_\theta W = \mathbf{E}_\theta\mathbf{E}_\theta(Y|Z) = \mathbf{E}_\theta Y$, so that W is also unbiased in Theorem 6.7.

Remark 6.11. What happens if Z is constant in the Rao-Blackwell Theorem? This seems desirable since then $W := \mathbf{E}_\theta(Y|Z)$ is also constant, so W has variance zero for any fixed $\theta \in \Theta$. But if g is not constant, then it is impossible for Z to be unbiased, hence W is not unbiased. Moreover, W is a function only of θ and not a function of the random sample. So, W is not a statistic.

Put another way, if Z does not have enough information, then conditioning on Z in the Rao-Blackwell Theorem seems undesirable. On the other hand, if Z has excess information (i.e. Z is not complete), then this might also lead to no improvement in the variance. For example, if Z is the vector of order statistics, then conditioning on Z does not change anything, i.e. $\mathbf{E}_\theta(Y|Z) = Y$, i.e. conditioning on Z does not improve the variance at all.

Example 6.12. Let $X_1, \ldots, X_n$ be a random sample of size n with unknown mean $\mu \in \mathbb{R}$. Suppose we want to construct an estimator for the mean using the Rao-Blackwell Theorem 6.7. Let $t\colon \mathbb{R}^n \to \mathbb{R}$ so that $t(x_1, \ldots, x_n) := x_1$ for all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Let $Y := t(X_1, \ldots, X_n) = X_1$. Note that Y is unbiased since $\mathbf{E}Y = \mathbf{E}X_1 = \mu$. By Exercise 5.24 (v) and (iv),

$$W := \mathbf{E}(X_1|X_1, \ldots, X_n) = \mathbf{E}(X_1|X_1) = X_1.$$

That is, conditioning on the whole sample does not change the statistic X_1 at all, even though the sample itself $(X_1, \ldots, X_n)$ is sufficient for μ. So, sometimes the Rao-Blackwell procedure may not be helpful.

Now, let's instead condition on $\sum_{i=1}^n X_i$. Since the random variables are i.i.d., for any $1 \le k < \ell \le n$, the joint distribution of $(X_k, \sum_{i=1}^n X_i)$ is equal to the joint distribution of $(X_\ell, \sum_{i=1}^n X_i)$. So, by the definition of conditional expectation in Exercise 5.24,

$$\mathbf{E}(X_k|\sum_{i=1}^n X_i) = \mathbf{E}(X_\ell|\sum_{i=1}^n X_i).$$

Therefore, by Exercise 5.24(iv)

$$W := \mathbf{E}(X_1|\sum_{i=1}^n X_i) = \frac{1}{n}\sum_{j=1}^n \mathbf{E}(X_j|\sum_{i=1}^n X_i) = \frac{1}{n}\mathbf{E}(\sum_{j=1}^n X_j|\sum_{i=1}^n X_i) = \frac{1}{n}\sum_{i=1}^n X_i.$$

So, in this case the Rao-Blackwell Theorem 6.7 did in fact substantially improve our estimator $Y = X_1$, since W has variance of order n^{-1}, while Y has constant variance.

From the above example, we see that the choice of the sufficient statistic Z in Theorem 6.7 can make a significant difference in the variance of the new estimator, and the choice of the unbiased estimator does not seem very important. The Example suggests that conditioning on too much "excess information" is not helpful, since conditioning on the whole sample made no improvement in the variance of the estimator. And indeed, the following Theorem says that a complete sufficient statistic is essentially the best thing to condition on in Theorem 6.7.

Theorem 6.13 (Lehmann-Scheffé). *Let Z be a complete sufficient statistic for $\{f_\theta\colon \theta \in \Theta\}$ and let Y be an unbiased estimator for $g(\theta)$. Define $W := \mathbf{E}_\theta(Y|Z)$. Assume that $\ell(\theta, y)$ is convex in y, for all $\theta \in \Theta$. Then W is UMRU for $g(\theta)$. If $\ell(\theta, y)$ is strictly convex in y for all $\theta \in \Theta$, then W is unique.*

In particular, W is the unique UMVU for $g(\theta)$.

Proof. W is unbiased by Remark 6.10. We first show that W does not depend on Y. Let Y' be another unbiased estimator for $g(\theta)$. We show that

$$\mathbf{E}_\theta(Y|Z) = \mathbf{E}_\theta(Y'|Z), \qquad \forall\, \theta \in \Theta. \qquad (*)$$

By Exercise 5.24(ii),

$$\mathbf{E}_\theta(\mathbf{E}_\theta(Y|Z) - \mathbf{E}_\theta(Y'|Z)) = \mathbf{E}_\theta(Y - Y') = g(\theta) - g(\theta) = 0, \qquad \forall\, \theta \in \Theta.$$

Therefore, $\mathbf{E}_\theta(Y|Z) = \mathbf{E}_\theta(Y'|Z)$ by completeness of Z. (By the definition of conditional expectation in Exercise 6.8, $\mathbf{E}_\theta(Y|Z)$ and $\mathbf{E}_\theta(Y|Z)$ are both functions of Z. Also, recall that these are not explicit functions of θ since Z is sufficient, using Exercise 6.9.)

Now by the Rao-Blackwell Theorem 6.7,

$$r(\theta, Y') \geq r(\theta, \mathbf{E}_\theta(Y'|Z)) \overset{(*)}{=} r(\theta, \mathbf{E}_\theta(Y|Z)) = r(\theta, W), \qquad \forall\, \theta \in \Theta.$$

□

Remark 6.14. Let $Z\colon \Omega \to \mathbb{R}^k$ be a complete sufficient statistic for $\{f_\theta\colon \theta \in \Theta\}$, and let $h\colon \mathbb{R}^k \to \mathbb{R}^m$. Let $g(\theta) := \mathbf{E}_\theta h(Z)$ for all $\theta \in \Theta$. Then $h(Z)$ is unbiased for $g(\theta)$. By Exercise 5.24(iii)

$$W := \mathbf{E}_\theta(h(Z)|Z) = h(Z)\mathbf{E}_\theta(1|Z) = h(Z).$$

So, by Theorem 6.13, $h(Z)$ is UMVU for $g(\theta)$.

So, one way to find a UMVU is to come up with a function of a complete sufficient statistic that is unbiased for a given function $g(\theta)$.

Here are some methods for finding a UMVU, as suggested by Theorem 6.13.

Suppose we have a complete sufficient statistic $Z\colon \Omega \to \mathbb{R}^k$ (recall it may not exist) and we want to estimate $g(\theta)$, $g\colon \Theta \to \mathbb{R}$.

(1) (Condition Method/Rao-Blackwell) Find an unbiased estimator Y for $g(\theta)$. Then $\mathbf{E}_\theta(Y|Z)$ is UMVU.
(2) Solve for an $h\colon \mathbb{R}^k \to \mathbb{R}$ such that $\mathbf{E}_\theta h(Z) = g(\theta)$.
(3) Somehow guess an $h\colon \mathbb{R}^k \to \mathbb{R}$ such that $\mathbf{E}_\theta h(Z) = g(\theta)$.

Example 6.15. Suppose we are sampling from the Gaussian distribution with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Recall from the Factorization Theorem 5.4 and Exercise 5.23 that $(\overline{X}, S^2)$ is complete sufficient for (μ, σ^2). So $\overline{X}$ is UMVU for μ (with fixed σ) by method (3) above, since $\overline{X}$ is a function of the complete sufficient statistic $Z = (\overline{X}, S^2)$, using $h(x, y) := x$ and $g(\mu, \sigma^2) := \mu$. Similarly, S^2 is UMVU for σ^2 (with fixed μ) by method (3) above, since S^2 is a function of the complete sufficient statistic $Z = (\overline{X}, S^2)$, using $h(x, y) := y$ and $g(\mu, \sigma^2) := \sigma^2$.

If we wanted a UMVU for μ^2, we can use method (3) above, noting that $\mathbf{E}\overline{X}^2 = \mu^2 + \sigma^2/n$,

$$\mathbf{E}[\overline{X}^2 - S^2/n] = \mu^2.$$

So $\overline{X}^2 - S^2/n$ is UMVU for μ^2 (with fixed σ) by method (3) above, since $\overline{X}^2 - S^2/n$ is a function of the complete sufficient statistic $Z = (\overline{X}, S^2)$, using $h(x, y) := x^2 - y/n$ and $g(\mu, \sigma^2) := \mu^2$.

Example 6.16. Let us illustrate method (2) above for a binomial random variable X with known parameter n and unknown parameter $0 < \theta < 1$. Suppose we want to estimate $g(\theta) := \theta(1-\theta)$ (this is the variance of a Bernoulli random variable with parameter $0 < \theta < 1$.) We want to find $h\colon \mathbb{R} \to \mathbb{R}$ such that

$$\theta(1-\theta) = \mathbf{E}_\theta h(X) = \sum_{j=0}^{n} h(j) \binom{n}{j} \theta^j (1-\theta)^{n-j}.$$

Letting $a := \theta/(1-\theta)$, so that $a(1-\theta) = \theta$, i.e. $\theta = a/(1+a)$ and $1-\theta = 1/(1+a)$, we want to find $h(j)$ such that

$$(1-\theta)^{-n}\mathbf{E}_\theta h(X) = \sum_{j=0}^{n} \binom{n}{j} h(j)a^j = \theta(1-\theta)^{1-n} = (1+a)^{-1}a(1+a)^{n-1} = a(1+a)^{n-2}$$
$$= a\sum_{j=0}^{n-2} \binom{n-2}{j} a^j = \sum_{j=1}^{n-1} \binom{n-2}{j-1} a^j.$$

Since this holds for all $0 < \theta < 1$, i.e. for all $a > 0$, both polynomials in a must have the same coefficients, so that $h(0) = h(n) = 0$, and $h(j)\binom{n}{j} = \binom{n-2}{j-1}$ for all $1 \le j \le n-1$. That is, for all $1 \le j \le n-1$,

$$h(j) = \binom{n-2}{j-1} \Big/ \binom{n}{j} = \frac{(n-2)!}{n!} \frac{(n-j)!j!}{(n-j-1)!(j-1)!} = \frac{(n-j)j}{n(n-1)}.$$

So, the UMVU for $\theta(1-\theta)$ is

$$\frac{X(n-X)}{n(n-1)}.$$

Example 6.17. Let's illustrate method (1). Suppose we have $n \ge 2$ independent samples $X_1, \dots, X_n$ from the Bernoulli distribution with unknown parameter $0 < \theta < 1$. From Example 3.15 and Exercise 5.23 (or Example 5.21), $Z := \sum_{i=1}^n X_i$ is complete and sufficient for θ. Also, $\frac{1}{n}\sum_{i=1}^n X_i$ is unbiased for θ, so $\frac{1}{n}\sum_{i=1}^n X_i$ is UMVU for θ.

Suppose we want to estimate θ^2. We use the unbiased estimate $Y := X_1X_2$ (noting that $\mathbf{E}_\theta Y = \mathbf{E}_\theta X_1 \mathbf{E}_\theta X_2 = \theta^2$, by independence.) The UMVU is then $\mathbf{E}(Y|Z)$. Let $2 \le z \le n$ be an integer. Note that $Y = 1$ when $X_1 = X_2 = 1$ and $Y = 0$ otherwise. So,

$$\mathbf{E}_\theta(Y|Z=z) = \mathbf{E}_\theta(1_{X_1=X_2=1}|Z=z) = \mathbf{P}_\theta(X_1 = X_2 = 1|Z=z)$$
$$= \mathbf{P}_\theta(X_1 = X_2 = 1|\sum_{i=1}^n X_i = z) = \frac{\mathbf{P}_\theta(X_1 = X_2 = 1, \sum_{i=1}^n X_i = z)}{\mathbf{P}_\theta(\sum_{i=1}^n X_i = z))}$$
$$= \frac{\mathbf{P}_\theta(X_1 = X_2 = 1, \sum_{i=3}^n X_i = z-2)}{\mathbf{P}_\theta(\sum_{i=1}^n X_i = z))} = \frac{\theta^2\binom{n-2}{z-2}\theta^{z-2}(1-\theta)^{n-z}}{\binom{n}{z}\theta^z(1-\theta)^{n-z}}$$
$$= \frac{1}{n(n-1)} \frac{(n-z)!z!}{(n-z)!(z-2)!} = \frac{z(z-1)}{n(n-1)}.$$

Additionally, $\mathbf{E}_\theta(Y|Z=z) = 0 = \frac{z(z-1)}{n(n-1)}$ for $z = 0$ and for $z = 1$. So,

$$\mathbf{E}_\theta(Y|Z=z) = \frac{z(z-1)}{n(n-1)}, \qquad \forall\, 0 \le z \le n.$$

So, the UMVU for θ^2 is

$$\mathbf{E}_\theta(Y|Z) = \frac{Z(Z-1)}{n(n-1)}.$$

The above procedures work well when we have a complete sufficient statistic, and these procedures do not work when we do not have a complete sufficient statistic. The following result attempts to deal with the case that the complete sufficient statistic does not exist, while we would still like to find the UMVU. Consider e.g. that we have a UMVU W_1 for $g_1(\theta)$

and we have a UMVU W_2 for $g_2(\theta)$. Does it follow that W_1+W_2 is UMVU for $g_1(\theta)+g_2(\theta)$? If the complete sufficient statistic exists, this follows immediately from Lehman-Scheffé's Theorem 6.13. It turns out the answer is also yes even when Theorem 6.13 does not apply (i.e. when we don't have a complete sufficient statistic). This follows from the following Theorem.

Theorem 6.18 (**Alternate Characterization of UMVU**). *Let $\{f_\theta\colon \theta\in\Theta\}$ be a family of distributions and let W be an unbiased estimator for $g(\theta)$. Let $L_2(\Omega)$ be the set of statistics with finite second moment. Then $W\in L_2(\Omega)$ is UMVU for $g(\theta)$ if and only if $\mathbf{E}_\theta(WU)=0$ $\forall\,\theta\in\Theta$, for all $U\in L_2(\Omega)$ that are unbiased estimators of 0.*

Proof. Assume W is UMVU for $g(\theta)$. Let U be an unbiased estimator of 0. Let $s\in\mathbb{R}$ and consider $W+sU$. Note that $W+sU$ is an unbiased estimator for $g(\theta)$. Since W is UMVU, we have

$$\mathrm{Var}_\theta(W)\le \mathrm{Var}_\theta(W+sU)=\mathrm{Var}_\theta W+2s\mathbf{E}_\theta(W-\mathbf{E}_\theta W)U+s^2\mathrm{Var}_\theta U.$$

The minimum value occurs at $s=0$ if and only if the derivative in s is zero at $s=0$, so that $0=\mathbf{E}_\theta(W-\mathbf{E}_\theta W)U=\mathbf{E}_\theta WU$. So, this reasoning can be reversed, since if Y is any unbiased estimator for $g(\theta)$, then $U:=Y-W$ is an unbiased estimator for 0, and $Y=W+sU$ with $s=1$, so

$$\mathrm{Var}_\theta(Y)=\mathrm{Var}_\theta(U+W)=\mathrm{Var}_\theta U+\mathrm{Var}_\theta W\ge\mathrm{Var}_\theta W.$$

□

6.3. **Efficiency of an Estimator.** Another desirable property of an estimator is high efficiency. That is, the estimator is good with a small number of samples. One way to quantify "good" in the previous sentence is to define a notion of information and to try to maximize the information content of the estimator.

Definition 6.19 (**Fisher Information**). Let $\{f_\theta\colon\theta\in\Theta\}$ be a family of multivariable probability densities or probability mass functions. Assume $\Theta\subseteq\mathbb{R}$. Let X be a random vector with distribution f_θ. Define the **Fisher information** of the family to be

$$I(\theta)=I_X(\theta):=\mathbf{E}_\theta\Big(\frac{d}{d\theta}\log f_\theta(X)\Big)^2,\qquad\forall\,\theta\in\Theta,$$

if this quantity exists and is finite.

In order for the Fisher information to be well defined, the set $\{x\in\mathbb{R}^n\colon f_\theta(x)>0\}$ should not depend on θ, otherwise the derivative $\frac{d}{d\theta}\log f_\theta(X)$ might not be well-defined.

If $\{f_\theta\colon\theta\in\Theta\}$ are n-dimensional probability densities, note that

$$\mathbf{E}_\theta\frac{d}{d\theta}\log f_\theta(X)=\int_{\mathbb{R}^n}\frac{\frac{d}{d\theta}f_\theta(x)}{f_\theta(x)}f_\theta(x)dx=\int_{\mathbb{R}^n}\frac{d}{d\theta}f_\theta(x)dx=\frac{d}{d\theta}\int_{\mathbb{R}^n}f_\theta(x)dx=\frac{d}{d\theta}(1)=0.$$

Similarly, if $\{f_\theta\colon\theta\in\Theta\}$ are multivariable probability mass functions, $\mathbf{E}_\theta\frac{d}{d\theta}\log f_\theta(X)=0$. So, we could equivalently define

$$I(\theta)=\mathrm{Var}_\theta\Big(\frac{d}{d\theta}\log f_\theta(X)\Big),\qquad\forall\,\theta\in\Theta.$$

(Differentiation under the integral sign can be justified whenever Proposition 9.8 applies.) We also have another equivalent definition:

$$\begin{aligned}\mathbf{E}_\theta \frac{d^2}{d\theta^2} \log f_\theta(X) &= \int_{\mathbb{R}^n} \frac{d}{d\theta} \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) dx = \int_{\mathbb{R}^n} \frac{f_\theta(x) \frac{d^2}{d\theta^2} f_\theta(x) - \left(\frac{d}{d\theta} f_\theta(x)\right)^2}{[f_\theta(x)]^2} f_\theta(x) dx \\ &= \int_{\mathbb{R}^n} \frac{d^2}{d\theta^2} f_\theta(x) - \left(\frac{d}{d\theta} \log f_\theta(x)\right)^2 f_\theta(x) dx = 0 - I_X(\theta) = -I_X(\theta).\end{aligned}$$

The Fisher information expresses the amount of "information" a random variable has.

Example 6.20. Let $\sigma > 0$ and let $f_\theta(x) := \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\theta)^2/[2\sigma^2]}$ for all $\theta \in \Theta$, $x \in \mathbb{R}$. We have

$$I(\theta) = \mathrm{Var}_\theta\Big(\frac{d}{d\theta} \frac{-(X-\theta)^2}{2\sigma^2}\Big) = \frac{1}{\sigma^4} \mathrm{Var}_\theta(X - \theta) = \frac{1}{\sigma^2}.$$

For the Gaussian case, we interpret "more information" as σ small, since then the variance is small, so more "information" is conveyed by a single sample than when σ is large. The Fisher information also agrees with our intuitive notion of information since the information of a joint distribution of independent random variables is equal to the sum of the separate informations.

Proposition 6.21. *Let X be a random variable with distribution from $\{f_\theta \colon \theta \in \Theta\}$ (densities or mass functions). Let Y be a random variable with distribution from $\{g_\theta \colon \theta \in \Theta\}$ (densities or mass functions). Assume that X and Y are independent. Then*

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta), \qquad \forall\, \theta \in \Theta.$$

Proof. Since X and Y are independent, (X, Y) has distribution from the multivariate density $f_\theta(X) g_\theta(Y)$. Also, $\frac{d}{d\theta} \log f_\theta(X)$ and $\frac{d}{d\theta} \log g_\theta(Y)$ are independent for any $\theta \in \Theta$, so

$$\begin{aligned}I_{(X,Y)}(\theta) &= \mathrm{Var}_\theta\Big(\frac{d}{d\theta} \log[f_\theta(X) g_\theta(Y)]\Big) = \mathrm{Var}_\theta\Big(\frac{d}{d\theta}[\log f_\theta(X) + \log g_\theta(Y)]\Big) \\ &= \mathrm{Var}_\theta\Big(\frac{d}{d\theta} \log f_\theta(X)\Big) + \mathrm{Var}_\theta\Big(\frac{d}{d\theta} \log g_\theta(X)\Big) = I_X(\theta) + I_Y(\theta).\end{aligned}$$

□

Exercise 6.22. Let X be a random variable with distribution from $\{f_\theta \colon \theta \in \Theta\}$ (densities or mass functions). Let Y be a random variable with distribution from $\{g_\theta \colon \theta \in \Theta\}$ (densities or mass functions). Show that

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_{Y|X=x}(\theta), \qquad \forall\, \theta \in \Theta,\, x \in \mathbb{R}.$$

(If X, Y are continuous random variables, recall that $Y|X$ has density $f_{X,Y}(x, y)/f_X(x)$ for any fixed x. And if X, Y are discrete random variables, recall that $Y|X$ has mass function $\mathbf{P}(X = x, Y = y)/\mathbf{P}(Y = y)$. And if)

Theorem 6.23 (Cramér-Rao/ Information Inequality). *Let $X \colon \Omega \to \mathbb{R}^n$ be a random variable with distribution from a family of multivariable probability densities or probability mass functions $\{f_\theta \colon \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$. Let $t \colon \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X)$ be statistic. For any $\theta \in \Theta$ let $g(\theta) := \mathbf{E}_\theta Y$. Then*

$$\mathrm{Var}_\theta(Y) \geq \frac{|g'(\theta)|^2}{I_X(\theta)}, \qquad \forall\, \theta \in \Theta.$$

In particular, if Y is unbiased for θ,

$$\mathrm{Var}_\theta(Y) \geq \frac{1}{I_X(\theta)}, \qquad \forall\, \theta \in \Theta.$$

Equality occurs for some $\theta \in \Theta$ only when $\frac{d}{d\theta} \log f_\theta(X)$ and $Y - \mathbf{E}_\theta Y$ are multiples of each other. Also, if $I_X(\theta) = 0$, then $g'(\theta) = 0$.

(Differentiation under the integral sign in the proof can be justified whenever Proposition 9.8 applies. Also, we assume that $\{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$ does not depend on θ, and for a.e. $x \in \mathbb{R}^n$, $(d/d\theta) f_\theta(x))$ exists and is finite.)

Proof. For any $\theta \in \Theta$ let $g(\theta) := \mathbf{E}_\theta Y$. We assume that X is continuous, the discrete case being similar. Using $\mathbf{E}_\theta \frac{d}{d\theta} \log f_\theta(X) = 0$ and Remark 1.63,

$$|g'(\theta)| = \left| \frac{d}{d\theta} \int_{\mathbb{R}^n} f_\theta(x) t(x) dx \right| = \left| \int_{\mathbb{R}^n} \frac{d}{d\theta} \log f_\theta(x) t(x) f_\theta(x) dx \right| = \left| \mathbf{E}_\theta \frac{d}{d\theta} \log f_\theta(X) t(X) \right|$$
$$= \left| \mathrm{Cov}_\theta\Big(\frac{d}{d\theta} \log f_\theta(X), t(X) \Big) \right| \leq \sqrt{ \mathrm{Var}_\theta\Big(\frac{d}{d\theta} \log f_\theta(X) \Big) \mathrm{Var}_\theta(t(X)) } = \sqrt{ I_X(\theta) \mathrm{Var}_\theta(t(X)) }.$$

The equality case follows from Remark 1.63 and the known equality case of the Cauchy-Schwarz Inequality (see Theorem 1.99). □

For a one-parameter family of distributions, the equality case of Theorem 6.23 gives a new way to find a UMVU that avoids any discussion of complete sufficient statistics. To find a UMVU, we look for affine functions of $\frac{d}{d\theta} \log f_\theta(X)$.

Example 6.24. Suppose $f_\theta(x) := \theta x^{\theta-1} 1_{0<x<1}$ for all $x \in \mathbb{R}, \theta > 0$. (This is a beta distribution with $\beta = 1$.) We have

$$\frac{d}{d\theta} \log f_\theta(x) = \frac{1}{\theta} + \log x, \qquad \forall\, 0 < x < 1.$$

A vector $X = (X_1, \ldots, X_n)$ of n independent samples from f_θ is distributed according to the product $\prod_{i=1}^n f_\theta(x_i)$, so that

$$\frac{d}{d\theta} \log \prod_{i=1}^n f_\theta(x_i) = \sum_{i=1}^n \Big(\frac{1}{\theta} + \log x_i \Big) = n \Big(\frac{1}{\theta} + \frac{1}{n} \log \prod_{i=1}^n x_i \Big), \qquad \forall\, 0 < x_i < 1,\ 1 \leq i \leq n.$$

Define

$$Y := -\frac{1}{n} \log \prod_{i=1}^n X_i$$

Since $\mathbf{E}_\theta \frac{d}{d\theta} \log \prod_{i=1}^n f_\theta(X_i) = 0$, as shown after Definition 6.19, we have $\mathbf{E}_\theta Y = 1/\theta$ for all $\theta > 0$. Note that $Y - \mathbf{E}_\theta Y$ is a multiple of $\frac{d}{d\theta} \log \prod_{i=1}^n f_\theta(X_i)$. By the equality case of Theorem 6.23, Y must be UMVU for $1/\theta = \mathbf{E}_\theta Y$.

Theorem 6.23 suggests the following quantity represents the efficiency of an estimator.

Definition 6.25 (**Efficiency**). Let $X\colon \Omega \to \mathbb{R}^n$ be a random variable with distribution from a family of multivariable probability densities or probability mass functions $\{f_\theta\colon \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$. Let $t\colon \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X)$ be statistic. Define the **efficiency** of Y to be

$$\frac{1}{I_X(\theta)\mathrm{Var}_\theta(Y)}, \qquad \forall\, \theta \in \Theta,$$

if this quantity exists and is finite. If Z is another statistic, we define the **relative efficiency** of Y to Z to be

$$\frac{I_X(\theta)\mathrm{Var}_\theta(Z)}{I_X(\theta)\mathrm{Var}_\theta(Y)} = \frac{\mathrm{Var}_\theta(Z)}{\mathrm{Var}_\theta(Y)}, \qquad \forall\, \theta \in \Theta.$$

6.4. **Bayes Estimation.** In Bayes estimation, the unknown parameter $\theta \in \Theta$ is regarded instead as a random variable Ψ. The distribution of Ψ reflects our prior knowledge about the probable values of Ψ. Then, given that $\Psi = \theta$, the conditional distribution of $X|\Psi = \theta$ is assumed to be $\{f_\theta\colon \theta \in \Theta\}$, where $f_\theta\colon \mathbb{R}^n \to [0,\infty)$. Suppose $t\colon \mathbb{R}^n \to \mathbb{R}^k$, and we have a statistic $Y := t(X)$ and a loss function $\ell\colon \Theta \times \mathbb{R}^k \to \mathbb{R}$. Let $g\colon \Theta \to \mathbb{R}^k$.

Definition 6.26 (**Bayes Estimator**). A **Bayes estimator** Y for $g(\theta)$ with respect to Ψ is defined such that

$$\mathbf{E}\ell(g(\Psi), Y) \le \mathbf{E}\ell(g(\Psi), Z).$$

for all estimators Z. Here the expectation is with respect to both Ψ and Y.

Note that we have not made any assumptions about bias for Y or Z.
In order to find a Bayes estimator, it is sufficient to minimize the conditional risk.

Proposition 6.27. *Suppose there exists $t\colon \mathbb{R}^n \to \mathbb{R}^k$ such that, for almost every $x \in \mathbb{R}^n$, $Y := t(X)$ minimizes*

$$\mathbf{E}(\ell(g(\Psi), Z) \,|\, X = x),$$

over all estimators Z. Then $t(X)$ is a Bayes estimator for $g(\theta)$ with respect to Ψ.

Remark 6.28. The condition on almost every $x \in \mathbb{R}^n$ is with respect to the marginal, unconditional distribution of X given by

$$\mathbf{P}(X \in A) = \int_\Theta \mathbf{P}_\theta(X \in A) d\Psi(\theta), \qquad \forall\, A \subseteq \mathbb{R}^n.$$

We also emphasize that $t\colon \mathbb{R}^k \to \mathbb{R}^k$ can depend on the distribution of Ψ.

Proof. By assumption, $\mathbf{E}(\ell(g(\Psi), t(X)) \,|\, X = x) \le \mathbf{E}(\ell(g(\Psi), Y) \,|\, X = x)$ for any estimator Y, and for almost every x. Taking expected values of both sides, we then get $\mathbf{E}\ell(g(\Psi), t(X)) \le \mathbf{E}\ell(g(\Psi), Y)$. □

Example 6.29. Suppose $n = 1$, $g(\theta) = \theta$ and $\ell(\Psi, Y) = (\Psi - Y)^2$. The minimum value of

$$\begin{aligned}\mathbf{E}[(\Psi - t(X))^2 | X = x] &= \mathbf{E}[(\Psi - t(x))^2 | X = x] = \mathbf{E}[\Psi^2 - 2\Psi t(x) + (t(x))^2 | X = x] \\ &= \mathbf{E}[\Psi^2 | X = x] - 2t(x)\mathbf{E}(\Psi | X = x) + t(x)^2.\end{aligned}$$

occurs when $t(x) = \mathbf{E}(\Psi | X = x)$. So, define $t(x) := \mathbf{E}(\Psi | X = x)$ for any $x \in \mathbb{R}$. By Proposition 6.27, $t(X) = \mathbf{E}(\Psi | X)$ is the Bayes estimator for $g(\theta) := \theta$ with respect to Ψ.

Given that $\Psi = \theta > 0$, suppose X is uniform on the interval $[0, \theta]$. Also, assume that Ψ has the gamma distribution with parameters $\alpha = 2$ and $\beta = 1$, so that Ψ has density $\theta e^{-\theta} 1_{\theta>0}$. The joint distribution of X and Ψ is then

$$f_{\Psi,X}(\theta, x) = f_{X|\Psi=\theta}(x|\theta) f_\Psi(\theta) = \frac{1}{\theta} 1_{0<x<\theta} \theta e^{-\theta} 1_{\theta>0} = 1_{0<x<\theta} e^{-\theta}.$$

The marginal distribution of X is then

$$f_X(x) = \int_{-\infty}^{\infty} f_{\Psi,X}(\theta, x) d\theta = 1_{x>0} \int_x^{\infty} e^{-\theta} d\theta = e^{-x} 1_{x>0}.$$

So, the conditional distribution of Ψ given X is

$$f_{\Psi|X=x}(\theta|x) = \frac{f_{\Psi,X}(\theta, x)}{f_X(x)} = \frac{1_{0<x<\theta} e^{-\theta}}{e^{-x} 1_{x>0}} = 1_{0<x<\theta} e^{x-\theta}.$$

So,

$$\mathbf{E}(\Psi|X = x) = \int_{-\infty}^{\infty} \theta f_{\Psi|X=x}(\theta|x) d\theta = \int_x^{\infty} \theta e^{x-\theta} d\theta = e^x((x+1)e^{-x}) = x + 1.$$

So, the Bayes estimator minimizing mean squared error for this particular Ψ is $t(X) = X + 1$.

In contrast, the UMVU for θ for a single sample X is $2X$ by Theorem 6.13, since $2X$ is complete sufficient and unbiased for θ, and $\mathbf{E}_\theta(2X|2X) = 2X$. (For n samples, $(1 + 1/n)X_{(n)}$ is UMVU for θ.)

6.5. **Method of Moments.**

Definition 6.30 (**Consistency**)**.** Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of distributions. Let $Y_1, Y_2, \ldots$ be a sequence of estimators of $g(\theta)$. We say that $Y_1, Y_2, \ldots$ is **consistent** for $g(\theta)$ if, for any $\theta \in \Theta$, $Y_1, Y_2, \ldots$ converges in probability to the constant value $g(\theta)$, with respect to the probability distribution f_θ.

Typically, we will take Y_n to be a function of a random sample of size n, for all $n \geq 1$.

Example 6.31. Let $X_1, \ldots, X_n$ be a random sample of size n with distribution f_θ. The Weak Law of Large Numbers, Theorem 2.10, says that the sample mean is consistent when $\mathbf{E}_\theta |X_1| < \infty$ for all $\theta \in \Theta$. More generally, if $j \geq 1$ is a positive integer such that $\mathbf{E}_\theta |X_1|^j < \infty$ for all $\theta \in \Theta$, then the j^{th} sample moment

$$M_j = M_j(\theta) := \frac{1}{n} \sum_{i=1}^{n} X_i^j$$

is also consistent (as $n \to \infty$), i.e. as $n \to \infty$, M_j converges in probability to the j^{th} moment

$$\mu_j(\theta) := \mathbf{E} X_1^j.$$

Note also that if $h \colon \mathbb{R} \to \mathbb{R}$ is continuous, and if $Y_1, Y_2, \ldots$ is consistent for $g(\theta)$, then $h(Y_1), h(Y_2), \ldots$ is consistent for $h(g(\theta))$.

Definition 6.32 (**Method of Moments**)**.** Let $g \colon \Theta \to \mathbb{R}^k$. Suppose we want to estimate $g(\theta)$ for any $\theta \in \Theta$. Suppose there exists $h \colon \mathbb{R}^j \to \mathbb{R}^k$ such that

$$g(\theta) = h(\mu_1, \ldots, \mu_j).$$

Then the estimator

$$h(M_1, \dots, M_j)$$

is a **method of moments** estimator for $g(\theta)$.

Example 6.33. To estimate the mean μ, we can use $\Theta = \mathbb{R} = \{\mu_1 \in \mathbb{R}\}$, $j = 1$ and $h(\mu_1) = \mu_1$, so that a method of moments estimator of μ_1 is the sample mean M_1.

To estimate the standard deviation, we can use $\Theta = \mathbb{R} \times (0, \infty) = \{(\mu_1, \mu_2) \colon \mu_1 \in \mathbb{R}, \mu_2 > 0\}$, $j = 2$ and $h(\mu_1, \mu_2) = \sqrt{\mu_2 - \mu_1^2}$, so that a method of moments estimator of the standard deviation is $\sqrt{M_2 - M_1^2}$.

This approach is good in that it uses essentially no assumptions about model parameters. Perhaps for this reason, the method of moments is one of the oldest methods of point estimation, originating in the late 1800s. However, when information about model parameters is available, often the method of moments does not work well (despite being consistent).

Example 6.34. Suppose $X_1, \dots, X_n$ is a random sample of size n from the interval $[0, \theta]$ and $\theta > 0$ is unknown. As mentioned in Example 6.29, $(1 + 1/n)X_{(n)}$ is UMVU for θ. Since $\mathbf{E}_\theta X_1 = \theta/2$, a method of moment estimator for θ is $2M_1 = \frac{2}{n}\sum_{i=1}^n X_i$. This estimator is unbiased and consistent, but its variance is $\frac{1}{3n}\theta^2$, while the variance of the UMVU is

$$\begin{aligned}
&\frac{(n+1)^2}{n^2}\mathbf{E}X_{(n)}^2 - \theta^2 = \frac{(n+1)^2}{n^2}\int_0^\theta 2t\mathbf{P}(X_{(n)} > t)dt - \theta^2 \\
&= \theta^2\Big(\frac{(n+1)^2}{n^2} - 1\Big) - \frac{(n+1)^2}{n^2}\int_0^\theta 2t\mathbf{P}(X_{(n)} < t)dt \\
&= \theta^2\Big(\frac{(n+1)^2}{n^2} - 1\Big) - \frac{(n+1)^2}{n^2}\theta^{-n}\int_0^\theta 2tt^n dt = \theta^2\Big(\frac{(n+1)^2}{n^2} - 1\Big) - \frac{(n+1)^2}{n^2}\theta^2\frac{2}{n+2} \\
&= \frac{\theta^2}{n^2(n+2)}\Big((n+1)^2(n+2) - n^2(n+2) - 2(n+1)^2\Big) \\
&= \frac{\theta^2}{n^2(n+2)}\Big([(n+1)^2 - n^2](n+2) - 2(n+1)^2\Big) \\
&= \frac{\theta^2}{n^2(n+2)}\Big([2n+1](n+2) - 2(n+1)^2\Big) = \frac{\theta^2}{n^2(n+2)}(5n - 4n + 2 - 2) = \frac{\theta^2}{n(n+2)}.
\end{aligned}$$

Example 6.35. Suppose we have a binomial random variable with unknown parameters n, p. It is known that $\mathbf{E}X_1 = np$ and $\mathbf{E}X_1^2 = np(1-p) + n^2p^2$. So, we solve for n, p in the system of equations

$$M_1 = np, \qquad M_2 = np(1-p) + n^2p^2,$$

to get the estimator for n:

$$N := \frac{M_1^2}{M_1 - (M_2 - M_1^2)}$$

and the estimator for p:

$$P := \frac{M_1}{N}.$$

6.6. **Maximum Likelihood Estimator.** Let $X_1, \ldots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. So, we denote the joint distribution of $X_1, \ldots, X_n$ as

$$\prod_{i=1}^{n} f_\theta(x_i), \qquad \forall\, 1 \leq i \leq n.$$

If we have data $x \in \mathbb{R}^n$, recall that we defined the function $\ell \colon \Theta \to [0, \infty)$

$$\ell(\theta) := \prod_{i=1}^{n} f_\theta(x_i)$$

and called it the **likelihood function**.

Definition 6.36 (**Maximum Likelihood Estimator**)**.** The **maximum likelihood estimator** (MLE) Y is the estimator maximizing the likelihood function. That is, $Y := t(X)$, $t \colon \mathbb{R}^n \to \Theta$ and $t(x_1, \ldots, x_n)$ is defined to be any value of $\theta \in \Theta$ that maximizes the function

$$\prod_{i=1}^{n} f_\theta(x_i),$$

if this value of θ exists. A priori, the θ maximizing $\ell(\theta)$ might not exist, and it might not be unique

Remark 6.37. Maximizing the likelihood $\ell(\theta)$ is equivalent to maximizing $\log \ell(\theta)$, since log is monotone increasing.

It is relatively easy to construct examples where the MLE is not unique.

Example 6.38. Let $f_\theta(x_1) := 1_{[\theta, \theta+1]}(x_1)$ for all $x_1, \theta \in \mathbb{R}$. Then, for all $x_1, \ldots, x_n, \theta \in \mathbb{R}$, we have

$$\prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} 1_{[\theta, \theta+1]}(x_i) = \prod_{i=1}^{n} 1_{x_i \in [\theta, \theta+1]}.$$

So, if $x_1 = \cdots = x_n = 0$, we have

$$\prod_{i=1}^{n} f_\theta(x_i) = 1_{0 \in [\theta, \theta+1]} = 1_{\theta \in [-1, 0]}.$$

That is, any value of $\theta \in [-1, 0]$ is a maximum of the likelihood function, i.e. there are infinitely many maxima of the likelihood function. This is certainly not desirable.

If the likelihood function is continuous and Θ is compact, then at least one maximum of the likelihood function must exist.

A common assumption of a probability density function is that it is logarithmically concave. We will describe how this condition guarantees the uniqueness of the MLE. For a proof of consistency of the MLE under certain assumptions, see the Keener book, Theorem 9.11.

Recall that $\phi \colon \mathbb{R}^n \to [-\infty, \infty]$ is convex if for any $x, y \in \mathbb{R}^n$ with $x \neq y$ and $\forall\, t \in (0, 1)$,

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y).$$

And $\phi \colon \mathbb{R}^n \to [-\infty, \infty]$ is strictly convex if this inequality is always a strict inequality. We also say ϕ is concave if $-\log \phi$ is convex, and ϕ is strictly concave if $-\log \phi$ is strictly convex.

Definition 6.39 (**Log-Concave**). We say that $\phi\colon \mathbb{R}^n \to [0,\infty)$ is **logarithmically concave** or **log concave** if $\log\phi$ is concave, i.e. $-\log\phi$ is convex.

For example, the function $\phi(x) = e^{-x^2}$, $x \in \mathbb{R}$, is log concave, since $\log\phi$ is concave. If we allow ϕ to take infinite values, then $1_{[-1,0]}$ is log-concave, so Example 6.38 shows that log-concavity still does not guarantee uniqueness of the maximum of the likelihood function. However, strict log-concavity does guarantee uniqueness.

Proposition 6.40. *Let $f_\theta\colon \mathbb{R} \to [0,\infty)$ be a family of probability density functions, where $\theta \in \Theta = \mathbb{R}^k$. Fix $x_1,\dots,x_n \in \mathbb{R}$. Assume that the function*

$$\theta \mapsto f_\theta(x_i)$$

is strictly log-concave, for every $1 \le i \le n$. Assume that Θ is a convex set (for any $a,b \in \Theta$ and for any $0 < t < 1$, $at + (1-t)b \in \Theta$). Then the likelihood function

$$\theta \mapsto \prod_{i=1}^n f_\theta(x_i)$$

has at most one maximum value.

Proof. The function $\theta \mapsto \log f_\theta(x_i)$ is strictly concave for all $1 \le i \le n$, so the function

$$\theta \mapsto \sum_{i=1}^n \log f_\theta(x_i) = \log \prod_{i=1}^n f_\theta(x_i)$$

is strictly concave by Exercise 6.43. From Exercise 6.41, this function has at most one global maximum. □

Exercise 6.41. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a convex function. Let $x \in \mathbb{R}^n$ be a local minimum of f. Show that x is in fact a global minimum of f.

Show also that if f is strictly convex, then there is at most one global minimum of f.

Now suppose additionally that f is a C^1 function (all derivatives of f exist and are continuous), and $x \in \mathbb{R}^n$ satisfies $\nabla f(x) = 0$. Show that x is a global minimum of f.

Exercise 6.42. Let A be a real $m \times n$ matrix. Let $x \in \mathbb{R}^n$ and let $b \in \mathbb{R}^m$. Show that the function $f\colon \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \frac{1}{2}\|Ax - b\|^2$ is convex. Moreover, show that

$$\nabla f(x) = A^T(Ax - b), \qquad D^2 f(x) = A^T A.$$

(Here $D^2 f$ denotes the matrix of second derivatives of f.)

So, if $\nabla f(x) = 0$, i.e. if $A^T Ax = A^T b$, then x is the global minimum of f. And if A has full rank, then $A^T A$ is invertible, so that $x = (A^T A)^{-1} A^T b$ is the global minimum of f.

Exercise 6.43. Let $f_1,\dots,f_n\colon \mathbb{R} \to \mathbb{R}$ be n strictly convex functions on $\mathbb{R}$. Define $g\colon \mathbb{R}^n \to \mathbb{R}$ by

$$g(x_1,\dots,x_n) := \sum_{i=1}^n f(x_i), \qquad \forall\,(x_1,\dots,x_n) \in \mathbb{R}^n.$$

Show that $g\colon \mathbb{R}^n \to \mathbb{R}$ is strictly convex.

Exercise 6.44. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be a C^1 function (all derivatives of f exist and are continuous). Suppose there exists $z \in \mathbb{R}$ such that, for any $x_1 \in \mathbb{R}$, we have

$$f(x_1, z) < f(x_1, x_2), \qquad \forall\, x_2 \neq z.$$

Assume also that the function

$$x_1 \mapsto f(x_1, z)$$

is strictly convex. Show that f has at most one global minimum.

Example 6.45. Consider a random sample from a Gaussian distribution with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$, so that $\theta = (\mu, \sigma)$. The value of θ maximizing

$$\log \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x_i - \mu)^2/[2\sigma^2]) = \sum_{i=1}^{n} -\log\sigma - \frac{1}{2}\log(2\pi) - \frac{(x_i-\mu)^2}{2\sigma^2}$$

can be found by differentiating in the two parameters. We have

$$\frac{\partial}{\partial\mu}\log\ell(\theta) = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2}, \qquad \frac{\partial}{\partial\sigma}\log\ell(\theta) = \sum_{i=1}^{n} -\sigma^{-1} + \sigma^{-3}(x_i-\mu)^2,$$

Setting both terms equal to zero, we get

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

This is the unique critical point of the function $\ell(\theta)$. It remains to show that this critical point is the global maximum of $\ell(\theta)$. It follows from Exercise 6.44 that, if $z \neq \frac{1}{n}\sum_{i=1}^{n} x_i$, then

$$\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 < \sum_{i=1}^{n}(x_i - z)^2.$$

Therefore, for any such $z \in \mathbb{R}$

$$\log\ell(\frac{1}{n}\sum_{i=1}^{n} x_i, \sigma) > \log\ell(z, \sigma).$$

So, we need only show that $\log\ell(\frac{1}{n}\sum_{i=1}^{n} x_i, \sigma)$ is maximized when $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$. Since

$$\frac{\partial}{\partial\sigma}\log\ell(\theta) = \sigma^{-3}\sum_{i=1}^{n} -\sigma^2 + (x_i-\mu)^2,$$

the function $\sigma \mapsto \log\ell(\mu, \sigma)$ is increasing, and then decreasing, so that the global maximum occurs at the unique critical point.

We already know the sample mean is UMVU for the mean, by e.g. Exercise 5.10. Let

$$Y = Y_n = Y_n(X_1, \ldots, X_n) := \frac{1}{n}\sum_{j=1}^{n}\left(X_j - \frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

We also know from Proposition 4.7 that Y is asymptotically unbiased for σ^2, i.e.

$$\lim_{n\to\infty}\frac{\mathbf{E}Y}{\sigma^2} = \lim_{n\to\infty}\frac{n-1}{n} = 1.$$

We will show that Y has asymptotically optimal variance without using the exponential family. If we fix $\mu \in \mathbb{R}$ and look at the information of the n-dimensional Gaussian X, we get by modifying Example 6.20 and using Proposition 6.21

$$\begin{aligned} I_X(\sigma) = nI_{X_1}(\sigma) &= n\mathrm{Var}_\sigma\Big(\frac{d}{d\sigma}\frac{-(X_1-\mu)^2}{2\sigma^2}\Big) = n\sigma^{-6}\mathrm{Var}_\sigma[(X_1-\mu)^2] \\ &= n\sigma^{-6}\mathbf{E}_\sigma((X_1-\mu)^4 - \sigma^4) = 2n\sigma^{-2}. \end{aligned}$$

By the Cramér-Rao Inequality, Theorem 6.23, with $g(\sigma) = \mathbf{E}_\sigma(Y) = \sigma^2(n-1)/n$ (using Proposition 4.7), the variance of any unbiased estimator Z of $\sigma^2(n-1)/n$ satisfies

$$\mathrm{Var}_\sigma(Z) \geq \frac{|g'(\sigma)|^2}{I_X(\sigma)} = \frac{4\sigma^2(n-1)^2}{n^2 2n\sigma^{-2}} = \frac{2\sigma^4(n-1)^2}{n^3}.$$

And by Proposition 4.7,

$$\mathrm{Var}_\sigma(Y) = \mathrm{Var}_\sigma\Big[\frac{\sigma^2}{n}\frac{1}{\sigma^2}\sum_{j=1}^n\Big(X_j - \frac{1}{n}\sum_{i=1}^n X_i\Big)^2\Big] = \frac{\sigma^4}{n^2}2(n-1) = \frac{2\sigma^4(n-1)}{n^2}.$$

In summary,

$$\lim_{n\to\infty}\frac{\mathbf{E}Y}{\sigma^2} = 1, \qquad \lim_{n\to\infty}\frac{\mathrm{Var}_\sigma(Y)}{|g'(\sigma)|^2/I_X(\sigma)} = 1.$$

That is, the estimator Y is asymptotically unbiased (as $n \to \infty$) and it asymptotically achieves the optimal variance bound in the Cramér-Rao Inequality.

Example 6.46. Consider a random sample that is uniform on $[0,\theta]$ with $\theta > 0$ unknown. The value of θ maximizing

$$\prod_{i=1}^n \frac{1}{\theta}1_{[0,\theta]}(x_i) = \theta^{-n}1_{x_1,\ldots,x_n\in[0,\theta]} = \theta^{-n}1_{x_{(1)},x_{(n)}\in[0,\theta]}$$

occurs when θ is as small as possible such that the likelihood is nonzero, since θ^{-n} is a decreasing function in θ. Once $\theta < x_{(n)}$, this expression is zero, so the smallest value of θ giving a nonzero likelihood is $\theta = x_{(n)}$. So, the unique global maximum occurs at $\theta = x_{(n)}$, so that $X_{(n)}$ is the MLE for θ. In contrast, recall that the UMVU for θ is $(1+1/n)X_{(n)}$, so both are asymptotically equivalent, though the MLE is biased.

Example 6.47. Consider a random sample from the exponential density $1_{x>0}\theta e^{-\theta x}$ with $\theta > 0$ unknown. Let $x_1,\ldots,x_n > 0$. Then

$$\log\prod_{i=1}^n \theta e^{-\theta x_i} = \log\theta - \theta\sum_{i=1}^n x_i.$$

So,

$$\frac{d}{d\theta}\log\prod_{i=1}^n \theta e^{-\theta x_i} = \frac{n}{\theta} - \sum_{i=1}^n x_i.$$

As a function of θ, the likelihood is increasing for small θ and decreasing for large θ, so there is a unique maximum of

$$Y := \frac{1}{\frac{1}{n}\sum_{i=1}^n X_i},$$

which is the MLE for θ.

To find the asymptotic efficiency of the MLE, recall that the exponential distribution has mean θ^{-1} and variance θ^{-2}, so by the Central Limit Theorem 2.13, $\sqrt{n}(\overline{X}_n - \theta^{-1})$ converges in distribution to a Gaussian random variable with mean 0 and variance θ^{-2} as $n \to \infty$. So, the Delta Method, Theorem 4.14, with $g(x) = 1/x$, $g'(x) = -1/x^2$ for all $x > 0$, shows that

$$\sqrt{n}\Big(\frac{1}{\overline{X}_n} - \theta\Big) = \sqrt{n}\Big(\frac{1}{\overline{X}_n} - g(1/\theta)\Big)$$

converges in distribution to a Gaussian random variable with mean 0 and variance $(g'(1/\theta))^2\theta^{-2} = \theta^2$ as $n \to \infty$. That is, (using also Theorem 4.16)

$$\operatorname{Var}(Y) = \operatorname{Var}\Big[n^{-1/2}\sqrt{n}\Big(\frac{1}{\overline{X}_n} - \theta\Big)\Big] = \frac{1}{n}\theta^2(1 + o(1)).$$

On the other hand, the information inequality, Theorem 6.23, says the smallest possible variance of an unbiased estimator of θ is

$$1/\operatorname{Var}\Big(\frac{n}{\theta} - \sum_{i=1}^{n} X_i\Big) = 1/(n\theta^{-2}) = \theta^2/n.$$

So, the MLE asymptotically achieves the optimal variance for an estimator of θ.

Example 6.48. Consider a random sample from the exponential density $1_{x>0}\theta e^{-\theta x}$ with $\theta > 0$ unknown. That is, we continue the previous example. Instead of finding an MLE for θ, suppose we want an MLE for $e^{-\theta}$. From the previous example, we can immediately conclude that

$$\psi = e^{-1/\sum_{i=1}^{n} x_i}.$$

by with $g(\theta) := e^{-\theta}$. Proposition 6.49.

Proposition 6.49 (**Functional Equivariance of MLE**). *Let $g\colon \Theta \to \Theta'$ be a bijection. Suppose Y is the MLE of θ. Then $g(Y)$ is the MLE of $g(\theta)$.*

Proof. By definition of the MLE Y, $Y(X_1, \dots, X_n)$ achieves the maximum value of $\theta \mapsto \ell(\theta)$. Writing $\ell(\theta) = \ell(g^{-1}g(\theta))$, we have the equivalent statement: $g(Y)(X_1, \dots, X_n)$ achieves the maximum value of $\theta' \mapsto \ell(g^{-1}(\theta'))$. □

So, unlike the UMVU, once we know the MLE for θ, we can easily get the MLE for invertible functions of θ.

Lemma 6.50 (**Likelihood Inequality**). *Let $X\colon \Omega \to \mathbb{R}^n$ be a random variable with probability density $f_\theta\colon \mathbb{R}^n \to [0, \infty)$. Let $f_\omega\colon \mathbb{R}^n \to [0, \infty)$ be another probability density. Assume that the probability laws $\mathbf{P}_\theta$ and $\mathbf{P}_\omega$ corresponding to f_θ and f_ω are not equal. Then the* ***Kullback-Leibler information***

$$I(\theta, \omega) := \mathbf{E}_\theta \log \frac{f_\theta(X)}{f_\omega(X)}$$

satisfies $I(\theta, \omega) > 0$.

Remark 6.51. If $\mathbf{P}_\theta(f_\omega(X) = 0 \text{ and } f_\theta(X) > 0) > 0$, then define $I(\theta, \omega) := \infty$, so there is nothing to prove. Also, in the definition of $I(\theta, \omega)$, if both densities take value zero, we define the ratio of zero over zero to be 1.

Proof. We may assume that $\mathbf{P}_\theta(f_\omega(X) = 0 \text{ and } f_\theta(X) > 0) = 0$. Note that $f_\theta(X) > 0$ with probability one with respect to $\mathbf{P}_\theta$. By Jensen's Inequality, Exercise 1.91,

$$-I(\theta,\omega) = \mathbf{E}_\theta \log \frac{f_\omega(X)}{f_\theta(X)} \leq \log \mathbf{E}_\theta \frac{f_\omega(X)}{f_\theta(X)} = \log \int_{x\in\mathbb{R}^n:\ f_\theta(x)>0} \frac{f_\omega(x)}{f_\theta(x)} f_\theta(x)dx \leq \log(1) = 0.$$

If $I(\theta,\omega) = 0$, then both of the inequalities above are equalities. The last inequality being an equality implies that $\{x \in \mathbb{R}^n\colon f_\theta(x) > 0\}$ and $\{x \in \mathbb{R}\colon f_\omega(x) > 0\}$ are equal almost everywhere. Since log is strictly concave, equality in the application of Jensen's Inequality implies that $\frac{f_\omega(X)}{f_\theta(X)}$ is constant almost surely (with respect to the probability law $\mathbf{P}_\theta$), therefore the densities f_ω and f_θ must be proportional, hence equal almost surely with respect to $\mathbf{P}_\theta$, so their corresponding probability laws are equal. □

Theorem 6.52 (Consistency of MLE). *Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}^k$ be i.i.d. random variables with common probability density $f_\theta\colon \mathbb{R}^k \to [0,\infty)$. Let $\Theta \subseteq \mathbb{R}^m$. Suppose Θ is compact and $f_\theta(x_1)$ is a continuous function of θ for a.e. $x_1 \in \mathbb{R}^k$. (Then a maximum of $\ell(\theta)$ exists, since it is a continuous function on a compact set.) Fix $\theta \in \Theta$. Assume that $\mathbf{E}_\theta \sup_{\theta'\in\Theta} |\log f_{\theta'}(X_1)| < \infty$, and $\mathbf{P}_\theta \neq \mathbf{P}_{\theta'}$, for all $\theta' \neq \theta$ with $\theta' \in \Theta$. Then, as $n \to \infty$, an MLE Y_n of θ converges in probability to the constant function θ, with respect to $\mathbf{P}_\theta$.*

Proof. For simplicity we assume that Θ is finite. For a full proof, see the Keener book, Theorem 9.11. Fix $\theta \in \Theta$.

For any $\theta' \in \Theta$ and $n \geq 1$, let $\ell_n(\theta') := \frac{1}{n}\sum_{i=1}^n \log f_{\theta'}(X_i)$. Denote $\Theta = \{\theta, \theta_1, \ldots, \theta_k\}$. By the Weak Law of Large Numbers, Theorem 2.10, for any $\theta' \in \Theta$, $\ell_n(\theta')$ converges in probability with respect to $\mathbf{P}_\theta$ to the constant $\mu(\theta') := \mathbf{E}_\theta \log f_{\theta'}(X_1)$ as $n \to \infty$. Since $\mathbf{P}_\theta \neq \mathbf{P}_{\theta'}$, for all $\theta' \neq \theta$, we have $\mu(\theta) > \mu(\theta')$ for all $\theta' \in \Theta$ with $\theta' \neq \theta$, by Lemma 6.50 (since $I(\theta,\theta') = \mu(\theta) - \mu(\theta') > 0$). For any $n \geq 1$, let

$$A_n := \{\ell_n(\theta) > \ell_n(\theta_j), \quad \forall\, 1 \leq j \leq k\}.$$

Then $\lim_{n\to\infty} \mathbf{P}_\theta(A_n) = 1$, and on the set A_n, the MLE Y_n is well-defined and unique with $Y_n = \theta$, so $\{Y_n = \theta\}^c \subseteq A_n^c$, and for any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}_\theta(|Y_n - \theta| > \varepsilon) \leq \lim_{n\to\infty} \mathbf{P}_\theta(A_n^c) = 0.$$

□

If $g\colon \Theta \to \Theta'$ is a continuous bijection, it follows from Proposition 6.49 that the MLE for $g(\theta)$ is also consistent, using also Proposition 2.36(ii) (for convergence in probability).

The above Theorem is analogous to a weak law of large numbers, since it gives convergence in probability of the MLE. Continuing this analogy, the following Theorem is analogous to the Central Limit Theorem, since it gives the limiting distribution of the MLE.

Theorem 6.53 (Limiting Distribution of MLE). *Let $\{f_\theta\colon \theta \in \Theta\}$ be a family of probability density functions, so that $f_\theta\colon \mathbb{R}^k \to [0,\infty)\ \forall\ \theta \in \Theta$. Let $X_1, X_2, \ldots$ be i.i.d. such that X_1 has density f_θ. Let $\Theta \subseteq \mathbb{R}$. Assume the following*

(i) *The set $A := \{x \in \mathbb{R}^k\colon f_\theta(x) > 0\}$ does not depend on θ.*
(ii) *For every $x \in A$, $\partial^2 f_\theta(x)/\partial\theta^2$ exists and is continuous in θ.*
(iii) *The Fisher Information $I_{X_1}(\theta)$ exists and is finite, with $\mathbf{E}_\theta \frac{d}{d\theta} \log f_\theta(X_1) = 0$ and*

$$I_{X_1}(\theta) = \mathbf{E}_\theta\left(\frac{d}{d\theta}\log f_\theta(X_1)\right)^2 = -\mathbf{E}_\theta \frac{d^2}{d\theta^2}\log f_\theta(X_1) > 0.$$

(iv) *For every θ in the interior of Θ, $\exists\ \varepsilon > 0$ such that*

$$\mathbf{E}_\theta \sup_{\theta' \in [\theta-\varepsilon, \theta+\varepsilon]} \left| \frac{d^2}{d[\theta']^2} \log f_{\theta'}(X_1) \right| < \infty.$$

(v) *An MLE Y_n of θ exists and is consistent.*

Then, for any θ in the interior of Θ, as $n \to \infty$,

$$\sqrt{n}(Y_n - \theta)$$

converges in distribution to a mean zero Gaussian with variance $\frac{1}{I_{X_1}(\theta)}$, with respect to $\mathbf{P}_\theta$.

Remark 6.54. Combining this Theorem with Proposition 6.49, under the above assumptions (and also if the variance of the MLE converges, i.e. we can apply something like Theorem 4.16), the MLE for θ achieves the asymptotically optimal variance in the Cramér-Rao Inequality, Theorem 6.23. The same holds for a twice continuously differentiable, invertible function of θ.

Proof. For simplicity we assume that Θ is finite. For a full proof, see the Keener book, Theorem 9.14. Fix $\theta \in \Theta$. (When Θ is finite, it has no interior, so the theorem is vacuous in this case, but the proof below is meant to illustrate the general case while avoiding a few technicalities.)

For any $\theta' \in \Theta$ and $n \geq 1$, let $\ell_n(\theta') := \frac{1}{n} \sum_{i=1}^n \log f_{\theta'}(X_i)$.

Choose $\varepsilon > 0$ sufficiently small such that $[\theta - \varepsilon, \theta + \varepsilon] \cap \Theta = \{\theta\}$. For any $n \geq 1$, let A_n be the event that $Y_n = \theta$. Since $Y_1, Y_2, \ldots$ is consistent by Assumption (v), $\lim_{n\to\infty} \mathbf{P}_\theta(A_n) = 1$. Since Y_n maximizes ℓ_n, we have $\ell_n'(Y_n) = 0$ on A_n. (Since Θ is finite, this is not true, so take it as an additional assumption.) Taylor expanding ℓ_n' then gives

$$0 = \ell_n'(Y_n) = \ell_n'(\theta) + \ell_n''(Z_n)(Y_n - \theta), \qquad \text{if } A_n \text{ occurs},$$

where Z_n lies between θ and Y_n. Rewriting this equation gives

$$\sqrt{n}(Y_n - \theta) = \frac{\sqrt{n}\ell_n'(\theta)}{-\ell_n''(Z_n)}, \qquad \text{if } A_n \text{ occurs and } \ell_n''(Z_n) \neq 0. \qquad (*)$$

By Assumption (iii), the summed terms in $\ell_n'(\theta)$ i.i.d. random variables with mean zero and variance $I_{X_1}(\theta)$. So, the Central Limit Theorem 2.13 says that $\sqrt{n}\ell_n'(\theta)$ converges in distribution to a mean zero Gaussian with variance $I_{X_1}(\theta)$.

We now examine the denominator of $(*)$. By Assumption (iv) and the Weak Law of Large Numbers, $\ell_n''(\theta')$ converges in probability to $\mathbf{E}_\theta \ell_n''(\theta')$. Since $|Z_n - \theta| \leq |Y_n - \theta|$ when A_n occurs, we conclude that Z_n also converges in probability to θ as $n \to \infty$. Since Z_n only takes finitely many values, $\ell_n''(Z_n)$ converges in probability to $\mathbf{E}_\theta \ell_n''(\theta) \overset{(iii)}{=} -I_{X_1}(\theta)$. So, $(*)$ implies that $\sqrt{n}(Y_n - \theta)$ converges in distribution as $n \to \infty$ to a mean zero Gaussian with variance

$$\frac{I_{X_1}(\theta)}{[I_{X_1}(\theta)]^2} = \frac{1}{I_{X_1}(\theta)}.$$

So, we are done by Exercise 6.55 with $B_n := A_n \cap \{\ell_n''(Z_n) \neq 0\}$ for all $n \geq 1$. □

Exercise 6.55. Suppose $W_1, W_2, \ldots$ are random variables that converge in distribution to a random variable W, and $U_1, U_2, \ldots$ is any sequence of random variables. Let $B_1, B_2, \ldots \subseteq \Omega$ satisfy $\lim_{n\to\infty} \mathbf{P}(B_n) = 1$. Then, as $n \to \infty$

$$W_n 1_{B_n} + U_n 1_{B_n^c}$$

converges in distribution to W.

6.7. **EM Algorithm.** Let $X\colon \Omega \to \mathbb{R}^n$ be a discrete or continuous random variable. Let $t\colon \mathbb{R}^n \to \mathbb{R}^m$ be a non-invertible function, and let $Y := t(X)$. For example, let $m < n$, and define t by $t(x_1, \ldots, x_n) := (x_1, \ldots, x_m)\ \forall\ (x_1, \ldots, x_n) \in \mathbb{R}^n$. Suppose we would ideally observe the sample X, but we can only observe the "incomplete" sample Y.

Suppose X has distribution from a family $\{f_\theta\colon \theta \in \Theta\}$ where $f_\theta\colon \mathbb{R}^n \to [0, \infty)$ for all $\theta \in \Theta$. To find the MLE of θ, we would ideally maximize

$$\log \ell(\theta) = \log f_\theta(X).$$

However, since X cannot be directly observed, we cannot compute $\ell(\theta)$ directly, so we might not be able to find the MLE. So, we instead approximate the maximum value of $\log \ell(\theta)$ by conditioning on Y.

The following algorithm tries to find the MLE for Y.

Algorithm 6.56 (**Expectation-Maximization (EM) Algorithm**)**.** Initalize $\theta_0 \in \Theta$. Fix $k \geq 1$. For all $1 \leq j \leq k$, repeat the following procedure:

- (**Expectation**) Given θ_{j-1}, let $\phi_j(\theta) := \mathbf{E}_{\theta_{j-1}}(\log f_\theta(X)|Y)$, for any $\theta \in \Theta$.
- (**Maximization**) Let $\theta_j \in \Theta$ achieve the maximum value of ϕ_j (if it exists).

Remark 6.57. In the case that Y is constant, each step of the algorithm is identical by the Likelihood Inequality, Lemma 6.50. In the case that $Y = X$, the algorithm just outputs the MLE of $Y = X$ in one step. In the case where $m < n$, $X_1, \ldots, X_n\colon \Omega \to \mathbb{R}$ are i.i.d. with common density $f_\theta\colon \mathbb{R} \to [0, \infty)$ and $t(x_1, \ldots, x_n) := (x_1, \ldots, x_m)\ \forall\ (x_1, \ldots, x_n) \in \mathbb{R}^n$,

$$\phi_j(\theta) := \mathbf{E}_{\theta_{j-1}}\Big(\sum_{i=1}^n \log f_\theta(X_i)\Big|(X_1, \ldots, X_m)\Big) = \sum_{i=1}^m \log f_\theta(X_i) + \mathbf{E}_{\theta_{j-1}} \sum_{i=m+1}^n \log f_\theta(X_i).$$

So, ϕ_j is the log likelihood for $Y = (X_1, \ldots, X_m)$, plus the expected value of the log likelihood for $X_{m+1}, \ldots, X_n$.

Note that we cannot apply the Likelihood Inequality 6.50 directly to ϕ_j, i.e. the maximum value of ϕ_j is not θ_{j-1}, in general.

Denote $f_{X|Y}(x|y)$ the conditional density (or conditional probability mass function) of X given $Y = y$.

Lemma 6.58. *Suppose X has density f_θ and $Y := t(X)$ has density h_θ. We then denote $g_\theta(x|y) := f_{X|Y}(x|y)$. Then for any $\theta \in \Theta$,*

$$\log h_\theta(Y) - \log h_{\theta_{j-1}}(Y) \geq \phi_j(\theta) - \phi_j(\theta_{j-1}).$$

Equality holds only when $g_\theta(X|y) = g_{\theta_{j-1}}(X|y)$ almost surely with respect to $\mathbf{P}_{\theta_{j-1}}$ (for fixed y).

Proof. Since $f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y)$, we have

$$\log f_Y(y) = \log f_{X,Y}(x,y) - \log f_{X|Y}(x|y).$$

Since $Y = t(X)$, $f_{X,Y}(x,y) = f_X(x)1_{y=t(x)}$. That is, when $y = t(x)$, we have

$$\log f_Y(y) = \log f_X(x) - \log f_{X|Y}(x|y) = \log f_\theta(x) - \log f_{X|Y}(x|y).$$

Using our streamlined notation, we write instead

$$\log h_\theta(y) = \log f_\theta(x) - \log g_\theta(x|y).$$

Multiplying both sides by $h_{\theta_{j-1}}(x|y)$ and integrating in x, we get

$$\mathbf{E}_{\theta_{j-1}}\Big(\log h_\theta(Y)\Big|Y=y\Big) = \mathbf{E}_{\theta_{j-1}}\Big(\log f_\theta(X)\Big|Y=y\Big) - \mathbf{E}_{\theta_{j-1}}\Big(\log g_\theta(X|y)\Big|Y=y\Big)$$

Setting also $\theta = \theta_{j-1}$ and subtracting one equality from the other, we get

$$\begin{aligned}\log h_\theta(y) - \log h_{\theta_{j-1}}(y) = {}& \mathbf{E}_{\theta_{j-1}}\Big(\log f_\theta(X)\Big|Y=y\Big) - \mathbf{E}_{\theta_{j-1}}\Big(\log f_{\theta_{j-1}}(X)\Big|Y=y\Big)\\ & - \mathbf{E}_{\theta_{j-1}}\Big(\log g_\theta(X|y)\Big|Y=y\Big) + \mathbf{E}_{\theta_{j-1}}\Big(\log g_{\theta_{j-1}}(X|y)\Big|Y=y\Big)\end{aligned}$$

From the Likelihood Inequality, Lemma 6.50, the sum of the last two terms is nonnegative, and it is zero only when $\log g_\theta(X|y) = \log g_{\theta_{j-1}}(X|y)$ almost surely with respect to $\mathbf{P}_{\theta_{j-1}}$ (for fixed y). In summary,

$$\log h_\theta(Y) - \log h_{\theta_{j-1}}(Y) \geq \phi_j(\theta) - \phi_j(\theta_{j-1}).$$

□

Proposition 6.59 (**EM Algorithm Improvement**). *Let $\theta_0, \dots, \theta_k$ be an output of the EM Algorithm 6.56. Then for all $1 \leq j \leq k$,*

$$\log h_{\theta_j}(Y) \geq \log h_{\theta_{j-1}}(Y).$$

Proof. By the definition of θ_j in Algorithm 6.56, $\phi_j(\theta_j) \geq \phi_j(\theta_{j-1})$. So, Lemma 6.58 says

$$\log h_{\theta_j}(Y) - \log h_{\theta_{j-1}}(Y) \geq 0.$$

And equality occurs only when $g_{\theta_j}(X|y) = g_{\theta_{j-1}}(X|y)$ almost surely with respect to $\mathbf{P}_{\theta_{j-1}}$ (for fixed y), or when $\theta_j = \theta_{j-1}$. □

6.8. **Additional Comments.** The Cramér-Rao and Limiting Distribution for the MLE have analogous statements when Θ is a vector space.

Theorem 6.60 (**Multiparameter Cramér-Rao/ Information Inequality**). *Let $X\colon \Omega \to \mathbb{R}^n$ be a random variable with distribution from a family of multivariable probability densities or probability mass functions $\{f_\theta\colon \theta \in \Theta\}$. Assume that $\Theta \subseteq \mathbb{R}^m$ is an open set. We assume that $\{x \in \mathbb{R}^n\colon f_\theta(x) > 0\}$ does not depend on θ, and for a.e. $x \in \mathbb{R}^n$, and for all $1 \leq i \leq m$, $(\partial/\partial\theta_i)f_\theta(x)$ exists and is finite. Define the **Fisher information** of the family to be the $m \times m$ matrix $I(\theta) = I_X(\theta)$, so that if $1 \leq i, j \leq m$, the (i,j) entry of $I(\theta)$ is*

$$\mathrm{Cov}_\theta\Big(\frac{\partial}{\partial\theta_i}\log f_\theta(X), \frac{\partial}{\partial\theta_j}\log f_\theta(X)\Big) = \mathbf{E}_\theta\Big(\frac{\partial}{\partial\theta_i}\log f_\theta(X)\cdot\frac{\partial}{\partial\theta_j}\log f_\theta(X)\Big), \qquad \forall\,\theta\in\Theta,$$

and assume this quantity exists and is finite. Moreover, assume that $I(\theta)$ is an invertible matrix. (It is symmetric positive semidefinite by e.g. Exercise 2.35, but it might have a zero eigenvalue, a priori.)

Let $t\colon \mathbb{R}^n \to \mathbb{R}^m$ and let $Y := t(X)$ be statistic. For any $\theta \in \Theta$, let $g(\theta) := \mathbf{E}_\theta Y$ so that $g\colon \Theta \to \Theta$. Assume that all first order partial derivatives of g exist and are continuous. We assume that the assumptions of Proposition 9.8 hold, so that we can differentiate under the integral sign. Let $Dg(\theta)$ denote the matrix of first order partial derivatives of g, and let $\mathrm{Var}_\theta(Y)$ denote the covariance matrix of Y. Then

$$\mathrm{Var}_\theta(Y) \geq (Dg(\theta))^T [I_X(\theta)]^{-1} Dg(\theta), \qquad \forall\, \theta \in \Theta.$$

This is an inequality for symmetric matrices, i.e. for any column vector $v \in \mathbb{R}^m$, we have

$$v^T \mathrm{Var}_\theta(Y) v \geq v^T (Dg(\theta))^T [I_X(\theta)]^{-1} Dg(\theta) v, \qquad \forall\, \theta \in \Theta.$$

In particular, if Y is unbiased for θ,

$$\mathrm{Var}_\theta(Y) \geq [I_X(\theta)]^{-1}, \qquad \forall\, \theta \in \Theta.$$

Theorem 6.61 (Limiting Distribution of MLE). *Let $\{f_\theta\colon \theta \in \Theta\}$ be a family of probability density functions, so that $f_\theta\colon \mathbb{R}^n \to [0,\infty)$ $\forall\ \theta \in \Theta$. Let $X_1, X_2, \ldots$ be i.i.d. such that X_1 has density f_θ. Let $\Theta \subseteq \mathbb{R}^m$. Assume the following*

(i) *The set $A := \{x \in \mathbb{R}^n\colon f_\theta(x) > 0\}$ does not depend on θ.*

(ii) *For every $x \in A$, $\forall\ 1 \leq i, j \leq m$, $\frac{\partial^2 f_\theta(x)}{\partial\theta_i \partial\theta_j}$ exists and is continuous in θ.*

(iii) *The Fisher Information $I_{X_1}(\theta)$ exists and is finite, with $\mathbf{E}_\theta \nabla_\theta \log f_\theta(X_1) = 0$ and*

$$I_{X_1}(\theta) = \mathbf{E}_\theta\Big(\frac{\partial}{\partial\theta_i} \log f_\theta(X) \cdot \frac{\partial}{\partial\theta_j} \log f_\theta(X)\Big) = -\mathbf{E}_\theta D_\theta^2 \log f_\theta(X_1).$$

(D_θ^2 denotes the matrix of iterated second order derivatives in θ.) Moreover, assume that $I_{X_1}(\theta)$ is an invertible matrix.

(iv) *For every θ in the interior of Θ, $\forall\ 1 \leq i, j \leq m$, $\exists\ \varepsilon > 0$ such that*

$$\mathbf{E}_\theta \sup_{\theta' \in [\theta-\varepsilon, \theta+\varepsilon]} \left| \frac{\partial^2}{\partial\theta_i' \partial\theta_j'} \log f_{\theta'}(X_1) \right| < \infty.$$

(v) *The MLE Y_n of θ is consistent.*

Then, for any θ in the interior of Θ, as $n \to \infty$,

$$\sqrt{n}(Y_n - \theta)$$

converges in distribution to a mean zero Gaussian random vector with covariance matrix $[I_{X_1}(\theta)]^{-1}$, with respect to $\mathbf{P}_\theta$.

7. Resampling and Bias Reduction

The goal of bias reduction is to begin with an estimator and a random sample of fixed size n, and to find a way to reduce the bias of the estimator. We already know that conditioning as in the Rao-Blackwell Theorem 6.7 can allow us to reduce variance and maintain the bias of an estimator. Unfortunately, reducing the bias can sometimes increase the variance of the estimator. Recall that any random variable X can be written as

$$\mathbf{E}(X-\theta)^2 = \mathbf{E}(X - \mathbf{E}X + \mathbf{E}X - \theta)^2 = \mathbf{E}(X - \mathbf{E}X)^2 + (\mathbf{E}X - \theta)^2.$$

From this equality, we can intuitively assert that reducing the variance of an estimator could increase its bias, while reducing the bias of an estimator could increase its variance. This tradeoff is known as the bias-variance tradeoff.

A standard way to reduce bias is to resample from our random sample. In jackknife resampling, we consider the sample of size n with one sample removed, and then average the estimator over all n ways of removing one sample.

7.1. **Jackknife Resampling.**

Definition 7.1. Let $\Theta \subseteq \mathbb{R}$. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}^d$ be i.i.d random variables so that X_1 has distribution $f_\theta \colon \mathbb{R}^d \to [0, \infty)$, $\theta \in \Theta$. Let $Y_1, Y_2, \ldots$ be a sequence of estimators for θ so that for any $n \geq 1$, $Y_n = t_n(X_1, \ldots, X_n)$ for some $t_n \colon \mathbb{R}^{nd} \to \Theta$. For any $n \geq 1$, define the **jackknife estimator** of Y_n to be

$$Z_n := nY_n - \frac{n-1}{n} \sum_{i=1}^{n} t_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

The jackknife estimator reduces the bias of the original estimator, as we now show.

Proposition 7.2. *Assume that $Y_1, Y_2, \ldots$ are asymptotically unbiased, so that there exists $a, b \in \mathbb{R}$ such that*

$$\mathbf{E}Y_n = \theta + a/n + b/n^2 + O(1/n^3), \qquad \forall\, n \geq 1. \qquad (*)$$

Then

$$\mathbf{E}Z_n = \theta + O(1/n^2).$$

And if $b = 0$ and the $O(1/n^3)$ term is zero in $(*)$, then Z_n is unbiased.

Proof. Let $n \geq 1$. Then

$$\begin{aligned}
\mathbf{E}Z_n &\overset{(*)}{=} n\theta + a + \frac{b}{n} + O(1/n^2) - \frac{n-1}{n} \sum_{i=1}^{n} \mathbf{E}t_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\
&\overset{(*)}{=} n\theta + a + \frac{b}{n} + O(1/n^2) - \frac{n-1}{n} \sum_{i=1}^{n} (\theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O(1/n^3)) \\
&= \theta + \frac{b}{n} - \frac{b}{n-1} + O(1/n^2) = \theta + O(1/n^2).
\end{aligned}$$

□

Example 7.3. The jackknife estimator of the sample mean is the sample mean.

$$\begin{aligned}
&nY_n - \frac{n-1}{n} \sum_{i=1}^{n} t_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\
&\quad = \sum_{i=1}^{n} X_i - \frac{1}{n} \sum_{i=1}^{n} (X_1 + \cdots + X_{i-1} + X_{i+1} + \cdots + X_n) \\
&\quad = \sum_{i=1}^{n} X_i - \frac{n-1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad \forall\, n \geq 1.
\end{aligned}$$

Example 7.4. Let $X_1, \dots, X_n$ be i.i.d. Bernoulli random variables with parameter $0 < \theta < 1$. The MLE for θ is the sample mean, so by the Functional Equivariance Property of the MLE, Proposition 6.49, the MLE for θ^2 is

$$Y_n := \Big(\frac{1}{n}\sum_{i=1}^n X_i\Big)^2, \qquad \forall\, n \geq 1.$$

This estimator is biased, since

$$\mathbf{E}Y_n = \frac{1}{n^2}\Big(n\theta + n(n-1)\theta^2\Big) = \theta^2 + \frac{1}{n}(\theta - \theta^2), \qquad \forall\, n \geq 1.$$

By Proposition 7.2, the jackknife estimator

$$Z_n := n\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big)^2 - \frac{n-1}{n}\sum_{i=1}^n \Big(\frac{1}{n-1}\sum_{j\in\{1,\dots,n\}:\, j\neq i} X_j\Big)^2, \qquad \forall\, n \geq 1.$$

is an unbiased estimator of θ^2.

7.2. **Bootstrapping.**

Definition 7.5. Let $X_1, \dots, X_n$ be a random sample of size n. Let $m \geq 1$. We define the **bootstrap sample** $W_1, \dots, W_m$ as follows. Given $X_1, \dots, X_n$, let $W_1, \dots, W_m$ be a random sample of size m uniformly distributed in the values $\{X_1, \dots, X_n\}$.

We typically take m significantly larger than n.

For example, if we are given a sample of the form $\{3, 3, 5, 6\}$, then W_1 has probability $1/2$ of taking the value 3.

Remark 7.6. Note that $W_1, \dots, W_m$ are conditionally independent, by their definition. Although the original sample consists of independent random variables, the bootstrap sample does not. The easiest way to see this is to show that the covariance of W_1 and W_2 is nonzero. Indeed, using the conditional independence, we have

$$\begin{aligned}\mathbf{E}W_1W_2 &= \mathbf{E}\Big[\mathbf{E}(W_1W_2|X_1,\dots,X_n)\Big] = \mathbf{E}\Big[\mathbf{E}(W_1|X_1,\dots,X_n)\cdot\mathbf{E}(W_2|X_1,\dots,X_n)\Big]\\ &= \mathbf{E}\Big[\Big(\mathbf{E}(W_1|X_1,\dots,X_n)\Big)^2\Big] = \mathbf{E}\overline{X}^2.\end{aligned}$$

Meanwhile

$$\mathbf{E}(\overline{W}|X_1,\dots,X_n) = \frac{1}{n}\sum_{i=1}^n\Big(\frac{1}{n}\sum_{j=1}^n X_j\Big) = \overline{X}. \qquad (\ddagger)$$

So, the covariance of W_1 and W_2 is

$$\mathbf{E}(W_1 - \mathbf{E}W_1)(W_2 - \mathbf{E}W_2) = \mathbf{E}W_1W_2 - (\mathbf{E}W_1)(\mathbf{E}W_2) = \mathbf{E}\overline{X}^2 - (\mathbf{E}\overline{X})^2 = \mathrm{Var}\overline{X} = \frac{\mathrm{Var}(X_1)}{n}.$$

So, if X_1 is nonconstant, this covariance is nonzero.

Example 7.7. Suppose $\mu := \mathbf{E}X_1$, $\sigma := \sqrt{\mathrm{Var}(X_1)}$, and $\gamma := \mathbf{E}(X_1 - \mu)^3$. Suppose we want to estimate μ^3. The method of moments estimator for μ^3 is then $\overline{X}^3$. However, this

estimator is biased. We have

$$\begin{aligned}\mathbf{E}\overline{X}^3 &= \mathbf{E}(\overline{X}-\mu+\mu)^3 \\ &= \mu^3 + 3\mu^2\mathbf{E}(\overline{X}-\mu) + 3\mu\mathbf{E}(\overline{X}-\mu)^2 + \mathbf{E}(\overline{X}-\mu)^3 \\ &= \mu^3 + 3\mu\sigma^2/n + \gamma/n^2 = \mu^3 + O(1/n). \qquad (*)\end{aligned}$$

Here we used

$$\mathbf{E}(\overline{X}-\mu)^2 = \mathbf{E}\Big(\frac{1}{n}\sum_{i=1}^n (X_i-\mu)\Big)^2 = \frac{1}{n^2}(n\mathbf{E}(X_1-\mu)^2 + n(n-1)\mathbf{E}(X_1-\mu)(X_2-\mu)) = \frac{n}{n^2}\sigma^2.$$

$$\begin{aligned}\mathbf{E}(\overline{X}-\mu)^3 &= \mathbf{E}\Big(\frac{1}{n}\sum_{i=1}^n (X_i-\mu)\Big)^3 = \frac{1}{n^3}\mathbf{E}\sum_{1\le i,j,k\le n}(X_i-\mu)(X_j-\mu)(X_k-\mu) \\ &= \frac{1}{n^3}\Big(n\mathbf{E}(X_1-\mu)^3 + \sum_{1\le i,j,k\le n:\ i\ne j\vee j\ne k\vee i\ne k}\mathbf{E}(X_i-\mu)(X_j-\mu)(X_k-\mu)\Big) = \frac{n}{n^3}\gamma.\end{aligned}$$

After conditioning on $X_1,\dots,X_n$, $\overline{Y}$ is the sample mean of i.i.d. uniform random variables in $\{X_1,\dots,X_n\}$, so after conditioning we can re-use formula $(*)$ with $\overline{Y}$ in place of $\overline{X}$, i.e.

$$\begin{aligned}&\mathbf{E}(\overline{Y}^3|X_1,\dots,X_n) \\ &= (\mathbf{E}(Y_1|X_1,\dots,X_n))^3 + 3(\mathbf{E}(Y_1|X_1,\dots,X_n))(\mathbf{E}((Y_1-\overline{X})^2|X_1,\dots,X_n))/n \\ &\qquad + \mathbf{E}((Y_1-\overline{X})^3|X_1,\dots,X_n)/n^2 \\ &= \overline{X}^3 + \frac{3}{n}\overline{X}\frac{1}{n}\sum_{i=1}^n\Big(X_i-\overline{X}\Big)^2 + \frac{1}{n}\sum_{i=1}^n\Big(X_i-\overline{X}\Big)^3/n^2.\end{aligned}$$

Here we used $\mathbf{E}(Y_1|X_1,\dots,X_n) = \overline{X}$ using the definition of Y_1, along with

$$\mathbf{E}((Y_1-\overline{X})^2|X_1,\dots,X_n) = \frac{1}{n}\sum_{i=1}^n\Big(X_i-\overline{X}\Big)^2, \quad \mathbf{E}((Y_1-\overline{X})^3|X_1,\dots,X_n) = \frac{1}{n}\sum_{i=1}^n\Big(X_i-\overline{X}\Big)^3.$$

The bias-reduced estimator of μ^3 is then defined to be the original estimator, minus the "conditional bias" of the bootstrap estimator:

$$\begin{aligned}&\overline{X}^3 - \Big(\mathbf{E}(\overline{Y}^3|X_1,\dots,X_n) - [\mathbf{E}(\overline{Y}|X_1,\dots,X_n)]^3\Big) \\ &\qquad = \overline{X}^3 - \frac{3\overline{X}\frac{1}{n}\sum_{i=1}^n\Big(X_i-\overline{X}\Big)^2}{n} - \frac{\frac{1}{n}\sum_{i=1}^n\Big(X_i-\overline{X}\Big)^3}{n^2}.\end{aligned}$$

This estimator has expected value

$$\mu^3 + \frac{3}{n^2}(\mu\sigma^2-\gamma) + \frac{6\gamma}{n^3} - \frac{2\gamma}{n^4} = \mu^3 + O(1/n^2). \qquad (**)$$

So, the bias is asymptotically better than $\overline{X}^3$. To justify $(**)$, note that $(*)$ with $n=1$ says $\mathbf{E}X_1^3 = \mu^3 + 3\mu\sigma^2 + \gamma$. Then, for any $1\le j\le n$, we have

$$\mathbf{E}X_j\sum_{i=1}^n X_i^2 = \mathbf{E}X_j^3 + \sum_{i\ne j}\mathbf{E}X_j\mathbf{E}X_i^2 = \mu^3 + 3\mu\sigma^2 + \gamma + (n-1)\mu(\sigma^2+\mu^2).$$

Summing over j, we similarly have

$$\mathbf{E}\overline{X}\sum_{i=1}^{n} X_i^2 = \mu^3 + 3\mu\sigma^2 + \gamma + (n-1)\mu(\sigma^2+\mu^2).$$

Therefore,

$$\begin{aligned}
&\mathbf{E}\overline{X}\frac{1}{n}\sum_{i=1}^{n}\Big(X_i - \overline{X}\Big)^2 = \mathbf{E}\overline{X}\Big(-\overline{X}^2 + \frac{1}{n}\sum_{i=1}^{n} X_i^2\Big)\\
&\overset{(*)}{=} -(\mu^3 + 3\mu\sigma^2/n + \gamma/n^2) + \mu^3/n + 3\mu\sigma^2/n + \gamma/n + (1-1/n)\mu(\sigma^2+\mu^2)\\
&= -\gamma/n^2 + \gamma/n + \mu\sigma^2 - \mu\sigma^2/n\\
&= \mu\sigma^2 + (\gamma - \mu\sigma^2)/n - \gamma/n^2.
\end{aligned}$$

To compute the remaining term in the expected value before $(**)$, we use

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}(X_i - \overline{X})^3 = \mathbf{E}(X_1 - \overline{X})^3 = \mathbf{E}(X_1 - \mu + \mu - \overline{X})^3\\
&\quad= \mathbf{E}(X_1-\mu)^3 + 3\mathbf{E}(X_1-\mu)^2(\mu-\overline{X}) + 3\mathbf{E}(X_1-\mu)(\mu-\overline{X})^2 + \mathbf{E}(\mu-\overline{X})^3\\
&\quad= \gamma - \frac{3}{n}\mathbf{E}(X_1-\mu)^3 + 3\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{E}(X_1-\mu)(\mu-X_i)^2 - \frac{\gamma}{n^2}\\
&\quad= \gamma - \frac{3}{n}\gamma + 3\frac{1}{n^2}\mathbf{E}(X_1-\mu)^3 - \frac{\gamma}{n^2}\\
&\quad= \gamma(1 - 3/n + 2/n^2)
\end{aligned}$$

In total, the expected value we get then agrees with the formula from $(**)$, using $(*)$:

$$\begin{aligned}
&\mu^3 + 3\mu\sigma^2/n + \gamma/n^2 - \frac{3}{n}(\mu\sigma^2 + (\gamma-\mu\sigma^2)/n - \gamma/n^2) - \frac{1}{n^2}\gamma(1 - 3/n + 2/n^2)\\
&= \mu^3 + \frac{1}{n^2}(\gamma - 3\gamma + 3\mu\sigma^2 - \gamma) + \frac{1}{n^3}(3\gamma + 3\gamma) - \frac{1}{n^4}2\gamma\\
&= \mu^3 + \frac{3}{n^2}(-\gamma + \mu\sigma^2) + \frac{6}{n^3}\gamma - \frac{2}{n^4}\gamma
\end{aligned}$$

8. Some Concentration of Measure

8.1. **Concentration for Independent Sums.** In certain cases, we can make rather strong conclusions about the distribution of sums of i.i.d. random variables, improving upon the laws of large numbers.

Theorem 8.1 (**Hoeffding Inequality/ Large Deviation Estimate**). *Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = -1) = 1/2$. Let $a_1, a_2, \ldots \in \mathbb{R}$. Then, for any $n \geq 1$,*

$$\mathbf{P}\Big(\sum_{i=1}^{n} a_i X_i \geq t\Big) \leq e^{-\frac{t^2}{2\sum_{i=1}^{n} a_i^2}}, \qquad \forall\, t \geq 0.$$

Consequently,

$$\mathbf{P}\Big(\Big|\sum_{i=1}^n a_i X_i\Big| \geq t\Big) \leq 2e^{-\frac{t^2}{2\sum_{i=1}^n a_i^2}}, \qquad \forall\, t \geq 0.$$

Proof. By dividing $a_1, \ldots, a_n$ by a constant, we may assume $\sum_{i=1}^n a_i^2 = 1$. Let $\alpha > 0$. Using the (exponential) moment method as in Markov's inequality, Corollary 1.93, and $\alpha t \geq 0$,

$$\mathbf{P}(\sum_{i=1}^n a_i X_i \geq t) = \mathbf{P}(e^{\alpha \sum_{i=1}^n a_i X_i} \geq e^{\alpha t}) \leq e^{-\alpha t}\mathbf{E}e^{\alpha \sum_{i=1}^n a_i X_i} = e^{-\alpha t}\prod_{i=1}^n \mathbf{E}e^{\alpha a_i X_i}.$$

The last equality used independence of $X_1, X_2, \ldots$ and Proposition 1.13. Using an explicit computation and Exercise 8.2,

$$\mathbf{E}e^{\alpha a_i X_i} = (1/2)(e^{\alpha a_i} + e^{-\alpha a_i}) = \cosh(\alpha a_i) \leq e^{\alpha^2 a_i^2/2}, \qquad \forall\, i \geq 1.$$

In summary, for any $t \geq 0$

$$\mathbf{P}(\sum_{i=1}^n a_i X_i \geq t) \leq e^{-\alpha t} e^{\alpha^2 \sum_{i=1}^n a_i^2/2} = e^{-\alpha t + \alpha^2/2}.$$

Since $\alpha > 0$ is arbitrary, we choose α to minimize the right side. This minimum occurs when $\alpha = t$, so that $-\alpha t + \alpha^2/2 = -t^2/2$, giving the first desired bound. The final bound follows by writing $\mathbf{P}(|\sum_{i=1}^n a_i X_i| \geq t) = \mathbf{P}(\sum_{i=1}^n a_i X_i \geq t) + \mathbf{P}(-\sum_{i=1}^n a_i X_i \geq t)$ and then applying the first inequality twice. □

Exercise 8.2. Show that $\cosh(x) \leq e^{x^2/2}$, $\forall\, x \in \mathbb{R}$.

In particular, Hoeffding's inequality implies that

$$\mathbf{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n X_i\Big| \geq t\Big) \leq 2e^{-nt^2/2}, \qquad \forall\, t \geq 0.$$

This inequality is much stronger than either Markov's or Cheyshev's inequality, since they only respectively imply that

$$\mathbf{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n X_i\Big| \geq t\Big) \leq \frac{1}{t}, \quad \mathbf{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n X_i\Big| \geq t\Big) \leq \frac{1}{nt^2}, \qquad \forall\, t \geq 0.$$

Note also that Hoeffding's inequality gives a quantitative bound for any fixed $n \geq 1$, unlike the (non-quantitative) limit theorems which only hold as $n \to \infty$.

Exercise 8.3 (Chernoff Inequality). Let $0 < p < 1$. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mathbf{P}(X_1 = 1) = p$ and $\mathbf{P}(X_1 = 0) = 1 - p$ for any $i \geq 1$. Then for any $n \geq 1$

$$\mathbf{P}\Big(\frac{1}{n}\sum_{i=1}^n X_i \geq t\Big) \leq e^{-np}\Big(\frac{ep}{t}\Big)^{tn}, \qquad \forall\, t \geq p.$$

Prove the same estimate for $\mathbf{P}(\frac{1}{n}\sum_{i=1}^n X_i \leq t)$ for any $t \leq p$. (Hint: $1 + x \leq e^x$ for any $x \in \mathbb{R}$, so $1 + (e^\alpha - 1)p \leq e^{(e^\alpha - 1)p}$.)

Exercise 8.4. For any natural number n and a parameter $0 < p < 1$, define an Erdös-Renyi graph on n vertices with parameter p to be a random graph (V, E) on a (deterministic) vertex set V of n vertices (thus (V, E) is a random variable taking values in the discrete space of all $2^{\binom{n}{2}}$ possible undirected graphs one can place on V) such that the events $\{i, j\} \in E$ for unordered pairs with $i, j \in V$ are independent and each occur with probability p.

Suppose we have an Erdös-Renyi random graph $G = (V, E)$ on n vertices with parameter $0 < p < 1$. Define $d := p(n-1)$.

- Show that d is the expected degree of each vertex in G. (The degree of a vertex $v \in V$ is the number of vertices connected to v by an edge in E.)
- Show that there exists a constant $c > 0$ such that the following holds. Assume $p \geq \frac{c \log n}{n}$. Then with probability larger than .9, all vertices of G have degrees in the range $(.9d, 1.1d)$. (Hint: first consider a single vertex, then use the union bound over all vertices.)

8.2. **Concentration for Lipschitz Functions.** One way to phrase the general question in the subject of concentration of measure is: how far is a random variable from its mean value? Hoeffding's Inequality says that linear functions of mean zero ± 1 valued independent random variables are exponentially close to their mean value. A similar statement can be made for bounded random variables (see Theorem 8.7 below). In order to answer the general question, we next consider Lipschitz functions of i.i.d. random variables. We focus on the Gaussian setting for simplicity.

For any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we denote $\|x\| := (x_1^2 + \cdots + x_n^2)^{1/2}$.

Theorem 8.5 (**Concentration of measure for Gaussians, Lipschitz function form**). *Let $f \colon \mathbb{R}^n \to \mathbb{R}$. Suppose that for all $x, y \in \mathbb{R}^n$, $|f(x) - f(y)| \leq \|x - y\|$, so that f is 1-Lipschitz. Let $X = (X_1, \ldots, X_n)$ be a mean zero Gaussian random vector with identity convariance matrix. Then for all $t > 0$,*

$$\mathbf{P}\left(x \in \mathbb{R}^n \colon \ |f(x) - \mathbf{E}f(X)| \geq t\right) \leq 2e^{-2t^2/\pi^2}.$$

Proof. We assume that f all partial derivatives of f exist and are continuous. Let $Y = (Y_1, \ldots, Y_n)$ be another mean zero Gaussian random vector with identity convariance matrix, such that Y and X are independent. Let $0 \leq \theta \leq \pi/2$ and define

$$Z_\theta := X \sin\theta + Y \cos\theta.$$

By rotation invariance of a Gaussian random vector, Z_θ and $\frac{d}{d\theta} Z_\theta = X \cos\theta - Y \sin\theta$ have the same joint distribution as X and Y (since the vectors $(\sin\theta, \cos\theta)$ and $(\cos\theta, -\sin\theta)$ are orthogonal in $\mathbb{R}^2$.) Let $\phi \colon \mathbb{R} \to [0, \infty)$ be a convex function. Using then Jensen's Inequality,

Exercise 1.91, then the Chain Rule, then Jensen's inequality and Fubini's Theorem,

$$\begin{aligned}
\mathbf{E}\phi(f(X)-\mathbf{E}f(Y)) &\le \mathbf{E}\phi(f(X)-f(Y)) = \mathbf{E}\phi\Big(\int_0^{\pi/2}\frac{d}{d\theta}f(Z_\theta)d\theta\Big)\\
&= \mathbf{E}\phi\Big(\int_0^{\pi/2}\langle(\nabla f)(Z_\theta),\frac{d}{d\theta}Z_\theta\rangle d\theta\Big) = \mathbf{E}\phi\Big(\frac{1}{\pi/2}\int_0^{\pi/2}\frac{\pi}{2}\langle(\nabla f)(Z_\theta),\frac{d}{d\theta}Z_\theta\rangle d\theta\Big)\\
&\le \mathbf{E}\frac{1}{\pi/2}\int_0^{\pi/2}\phi\Big(\frac{\pi}{2}\langle(\nabla f)(Z_\theta),\frac{d}{d\theta}Z_\theta\rangle\Big)d\theta = \frac{1}{\pi/2}\int_0^{\pi/2}\mathbf{E}\phi\Big(\frac{\pi}{2}\langle(\nabla f)(Z_\theta),\frac{d}{d\theta}Z_\theta\rangle\Big)d\theta\\
&= \frac{1}{\pi/2}\int_0^{\pi/2}\mathbf{E}\phi\Big(\frac{\pi}{2}\langle(\nabla f)(X),Y\rangle\Big)d\theta = \mathbf{E}\phi\Big(\frac{\pi}{2}\langle(\nabla f)(X),Y\rangle\Big)
\end{aligned}$$

Let $\alpha\in\mathbb{R}$ and let $\phi(x):=e^{\alpha x}$ for all $x\in\mathbb{R}$. Then using independence in Y and Fubini's Theorem,

$$\mathbf{E}\exp(\alpha[f(X)-\mathbf{E}f(Y)]) \le \mathbf{E}\exp\Big(\alpha\frac{\pi}{2}\sum_{i=1}^n\frac{\partial f}{\partial x_i}(X)Y_i\Big) = \mathbf{E}_X\prod_{i=1}^n\mathbf{E}_Y\exp\Big(\alpha\frac{\pi}{2}\frac{\partial f}{\partial x_i}(X)Y_i\Big).$$

Using an explicit computation, for any $s\in\mathbb{R}$ and for any $1\le i\le n$,

$$\mathbf{E}_Y e^{sY_i} = \int_{-\infty}^{\infty}e^{sy}e^{-y^2/2}\frac{dy}{\sqrt{2\pi}} = e^{s^2/2}\int_{-\infty}^{\infty}e^{-(y-s)^2/2}\frac{dy}{\sqrt{2\pi}} = e^{s^2/2}.$$

So, applying this inequality with $s=\alpha\frac{\pi}{2}\frac{\partial f}{\partial x_i}(X)$ for each $1\le i\le n$,

$$\mathbf{E}\exp(\alpha[f(X)-\mathbf{E}f(Y)]) \le \mathbf{E}\exp\Big(\alpha^2\frac{\pi^2}{8}\sum_{i=1}^n\Big(\frac{\partial f}{\partial x_i}(X)\Big)^2\Big) \le \exp\Big(\alpha^2\frac{\pi^2}{8}\Big).$$

(Since f is 1-Lipschitz, $|\langle\nabla f(x),y\rangle|\le 1$ for all $x,y\in\mathbb{R}^n$ with $\|y\|\le 1$. In particular, using $y:=\nabla f(x)/\|\nabla f(x)\|$, we get $\|\nabla f(x)\|\le 1$.) So,

$$\begin{aligned}
\mathbf{P}(f(X)-\mathbf{E}f(Y)>t) &= \mathbf{P}(\exp(\alpha[f(X)-\mathbf{E}f(Y)])>e^{\alpha t})\\
&\le e^{-\alpha t}\exp\Big(\alpha^2\frac{\pi^2}{8}\Big) = \exp\Big(-\alpha t+\alpha^2\frac{\pi^2}{8}\Big).
\end{aligned}$$

The minimum α occurs when $\alpha=4t/\pi^2$, so making this choice of α, we get

$$\mathbf{P}(f(X)-\mathbf{E}f(Y)>t)\le\exp(-2t^2/\pi^2).$$

Similarly, $\mathbf{P}(f(X)-\mathbf{E}f(Y)<-t)\le\exp(-2t^2/\pi^2)$, so that

$$\begin{aligned}
\mathbf{P}(|f(X)-\mathbf{E}f(Y)|>t) &= \mathbf{P}(f(X)-\mathbf{E}f(Y)>t)+\mathbf{P}(f(X)-\mathbf{E}f(Y)<-t)\\
&\le 2\exp(-2t^2/\pi^2).
\end{aligned}$$

□

Theorem 8.6 (Johnson-Lindenstrauss Lemma). *Let $x^{(1)},\dots,x^{(n)}\in\mathbb{R}^m$. Let $\varepsilon>0$. Then there exists a linear function $h\colon\mathbb{R}^m\to\mathbb{R}^{O(\varepsilon^{-2}\log n)}$ such that*

$$\big\|x^{(i)}-x^{(j)}\big\| \le \big\|h(x^{(i)})-h(x^{(j)})\big\| \le (1+\varepsilon)\big\|x^{(i)}-x^{(j)}\big\|, \qquad \forall\, 1\le i,j\le n.$$

One proves this via the probabilistic method. By concentration of measure, a random projection does what we require.

Proof. Fix $1 \le k \le m$. Let $\Pi\colon \mathbb{R}^m \to \mathbb{R}^m$ be the orthogonal projection such that

$$\Pi(z_1, \dots, z_m) := (z_1, \dots, z_k, 0, \dots, 0), \qquad \forall\, (z_1, \dots, z_m) \in \mathbb{R}^m.$$

Let $X = (X_1, \dots, X_m)$ be a standard m-dimensional Gaussian random vector. Define

$$a := \mathbf{E}\, \|\Pi X\|\,.$$

We will eventually show that $a \ge 10^{-2}\sqrt{k}$. Observe

$$\mathbf{E}\, \|\Pi X\|^2 = \mathbf{E} \sum_{i=1}^{k} X_i^2 = k\mathbf{E}X_1^2. = k. \qquad (*)$$

Now, we use Remark 1.38 and 8.5 for the 1-Lipschitz function $x \mapsto \|\Pi x\|$,

$$\begin{aligned}
\mathbf{E}\, \|\Pi X\|^4 &= \int_0^\infty 4u^3 \mathbf{P}(\|\Pi X\| \ge u) du \\
&= \int_0^{2a} 4u^3 \mathbf{P}(\|\Pi X\| \ge u) du + \int_{2a}^\infty 4u^3 \mathbf{P}(\|\Pi X\| \ge u) du \\
&\le \int_0^{2a} 4u^3 du + \int_{2a}^\infty 4u^3 \mathbf{P}(|\, \|\Pi X\| - a| > u/2) du \\
&\le 16a^4 + 8\int_{2a}^\infty u^3 e^{-u^2/2\pi^2} du = 16a^4 + 8(2\pi^2)(2a^2 + \pi^2) e^{-2a^2/\pi^2} \le 16a^4 + 2\pi^4 \\
&\le 16a^4 + 200k^2 \le 216 \left(\int_{\mathbb{R}^m} \|\Pi x\|^2 \gamma_m(x) dx \right)^2 \quad \text{, using Jensen's inequality and } (*).
\end{aligned}$$

So, if $Z := \|\Pi X\|$ is a random variable, we have shown that $\mathbf{E}Z^4 < c(\mathbf{E}Z^2)^2$ where $c := 216$. So, using Hölder's Inequality, Theorem 1.99, for $p = 3/2$, $q = 3$,

$$\mathbf{E}Z^2 = \mathbf{E}(Z^{2/3} Z^{4/3}) \le (\mathbf{E}Z)^{2/3} (\mathbf{E}Z^4)^{1/3} \le (\mathbf{E}Z)^{2/3} c^{1/3} (\mathbf{E}Z^2)^{2/3}.$$

Using this inequality and $(*)$,

$$\mathbf{E}Z \ge c^{-1/2} \sqrt{\mathbf{E}Z^2} \ge 216^{-1/2} \sqrt{k}. \qquad (**)$$

In summary, $a \ge 2^{-4}\sqrt{k}$ for a defined above.

Let A be an $m \times m$ matrix of i.i.d. standard Gaussian random variables. Fix $x^{(0)} \in \mathbb{R}^m$ with $\|x\| = 1$. By rotation invariance of the Gaussian measure, A and AQ have the same distribution where Q is a fixed $m \times m$ orthogonal matrix, so if we choose Q so that $Q(1, 0, \dots, 0)^T = x^{(0)}$, we get

$$\begin{aligned}
\mathbf{P}\left(A \in \mathbb{R}^{m\times m}\colon \left|\, \left\|\Pi A x^{(0)}\right\|_2 - a\right| \ge \varepsilon a\right) &= \mathbf{P}\left(A \in \mathbb{R}^{m\times m}\colon \left|\, \left\|\Pi A (1, 0, \dots, 0)^T\right\|_2 - a\right| \ge \varepsilon a\right) \\
&= \mathbf{P}\left(X \in \mathbb{R}^m \,|\, |\, \|\Pi X\| - a| \ge \varepsilon a\right).
\end{aligned}$$

So, by Theorem 8.5 applied to the 1-Lipschitz function $x \mapsto \|\Pi x\|$, and using $a \ge 2^{-4}\sqrt{k}$, for any $\varepsilon > 0$, and for any

$$\mathbf{P}\left(A \in \mathbb{R}^{m\times m}\colon \left|\, \left\|\Pi A x^{(0)}\right\|_2 - a\right| \ge \varepsilon a\right) \le 2e^{-2\varepsilon^2 a^2/\pi^2} \le 2e^{-2^{-10} k\varepsilon^2}.$$

Let $x^{(1)}, \dots, x^{(n)}$ be n points in $\mathbb{R}^m$. If $k \ge 2^{12}\varepsilon^{-2} \log n$, the union bound shows that

$$\mathbf{P}\left(A \in \mathbb{R}^{m\times m}\colon \exists\, i \ne j\colon \left| \left\| \Pi A \left(\frac{x^{(i)} - x^{(j)}}{\|x^{(i)} - x^{(j)}\|} \right) \right\| - a \right| \ge \varepsilon a \right) \le \binom{n}{2} 2e^{-2^{-10} k\varepsilon^2} < 1.$$

For any $1 \le i \le n$, define $y_i := \Pi A x^{(i)}/(a(1-\varepsilon))$. Then $\exists\ A \in \mathbb{R}^{n\times m}$ such that

$$1 \le \left\| \frac{y^{(i)} - y^{(j)}}{\|x^{(i)} - x^{(j)}\|} \right\| \le \frac{1+\varepsilon}{1-\varepsilon} \le 1 + 3\varepsilon, \qquad \forall\, 1 \le i, j \le n.$$

So, our required embedding is $h := \frac{\Pi A}{a(1-\varepsilon)}$, so that $h(x^{(i)}) = y^{(i)}$ for all $1 \le i \le n$. Note that h is linear and its nonzero entries form a rectangular matrix of i.i.d. Gaussians. Also, we can choose $k := \lceil 2^{12}\varepsilon^{-2} \log n \rceil$. (In fact, if we choose k to be slightly larger, then the probability becomes exponentially small, so essentially all A satisfies our desired property, hence essentially all linear projections $h\colon \mathbb{R}^n \to \mathbb{R}^{O(\varepsilon^{-2}\log n)}$ satisfy our desired property.) $\square$

8.3. **Additional Comments.** Hoeffding's inequality in Theorem 8.1 can be generalized to the following statement.

Theorem 8.7 (**Hoeffding Inequality/ Large Deviation Estimate**). *For all $i \ge 1$, let $a_i < b_i$ be real numbers. Let $X_1, X_2, \ldots$ be independent random variables with $\mathbf{P}(X_i \in [a_i, b_i]) = 1$. Then, for any $n \ge 1$,*

$$\mathbf{P}\Big(\sum_{i=1}^n X_i - \mathbf{E}\big(\sum_{j=1}^n X_j\big) \ge t \Big) \le e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}, \qquad \forall\, t \ge 0.$$

Consequently,

$$\mathbf{P}\Big(\Big| \sum_{i=1}^n X_i - \mathbf{E}\big(\sum_{j=1}^n X_j\big) \Big| \ge t \Big) \le 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}, \qquad \forall\, t \ge 0.$$

Lemma 8.8 (**Hoeffding's Lemma**). *Let $a < b$ be real numbers. Let X be a random variable with $\mathbf{P}(X \in [a,b]) = 1$. Then for any $\alpha \in \mathbb{R}$,*

$$\mathbf{E}e^{\alpha X} \le e^{\frac{1}{8}\alpha^2 (b-a)^2}.$$

Theorem 8.5 can be generalized to uniformly log-concave densities on Euclidean space (see Ledoux, "The Concentration of Measure Phenomenon," Proposition 2.18)

Theorem 8.9 (**Concentration of measure for Log-Concave Measures, Lipschitz function form**). *Let $f\colon \mathbb{R}^n \to \mathbb{R}$. Suppose that for all $x, y \in \mathbb{R}^n$, $|f(x) - f(y)| \le \|x - y\|$, so that f is 1-Lipschitz. Let $u\colon \mathbb{R}^n \to \mathbb{R}$ be a function such that $e^{-u(x)}$ is a probability density on $\mathbb{R}^n$. Assume there exists $c > 0$ such that the Hessian of u satisfies $\mathrm{Hess}(u)(x) \ge cI$, in the matrix sense. (That is, all eigenvalues of the Hessian of u are bounded below by c, for all $x \in \mathbb{R}^n$.) Let X have distribution e^{-u}. Then, for all $t > 0$,*

$$\mathbf{P}\left(x \in \mathbb{R}^n\colon\ |f(x) - \mathbf{E}f(X)| \ge t\right) \le 2e^{-ct^2/2}.$$

9. Appendix: Results from Analysis

Theorem 9.1. ***(Minkowski's Inequality)*** *Let $1 \le p \le \infty$, and let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be measurable. Then*

$$\left\| \int_{\mathbb{R}} f(x,y)dx \right\|_{p,dy} \le \int_{\mathbb{R}} \|f(x,y)\|_{p,dy}\, dx.$$

In particular, the integrand on the right is measurable, so if the right side is finite, then $\int_{\mathbb{R}} f(x,y)dx$ is defined for almost every $y \in \mathbb{R}$.

Proof. The right side is unchanged by replacing f with $|f|$, so without loss of generality we assume $f\colon \mathbb{R}^2 \to [0,\infty)$. The case $p = 1$ follows from Fubini's Theorem, Theorem 1.79. If $1 < p < \infty$, measurability follows from Fubini's Theorem, and the inequality follows from Fubini's Theorem and the Hölder inequality for y, Theorem 1.99 (for Lebesgue measure), with exponents p, p' (using $(p-1)p' = p$).

$$\begin{aligned}
\int_{\mathbb{R}} \left| \int_{\mathbb{R}} f(x,y)dx \right|^p dy &= \int_{\mathbb{R}} \left| \int_{\mathbb{R}} f(x,y)dx \right|^{p-1} \left| \int_{\mathbb{R}} f(x',y)dx' \right| dy \\
&= \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} f(x',y) \left| \int_{\mathbb{R}} f(x,y)dx \right|^{p-1} dy \Big) dx' \\
&\le \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} |f(x',y)|^p \, dy \Big)^{1/p} \Big(\int_{\mathbb{R}} | \int_{\mathbb{R}} f(x,y)dx|^{p'(p-1)} dy \Big)^{1/p'} dx' \\
&= \int_{\mathbb{R}} \|f(x',y)\|_{p,dy} \, dx' \cdot \Big(\int_{\mathbb{R}} | \int_{\mathbb{R}} f(x,y)dx|^p dy \Big)^{1/p'}.
\end{aligned}$$

If the right-most term is nonnegative and finite, we divide both sides by it to conclude, using $1 - 1/p' = 1/p$. If the right-most term is zero, there is nothing to prove. In the case that f is the indicator function of a rectangle, the right-most term is finite, so the Theorem holds in this case. The Monotone Convergence Theorem then implies that the Theorem holds for more general functions f.

The case $p = \infty$ takes more work. Measurability follows by approximating f by simple functions, and using that the limit of measurable functions is measurable. We then use duality. Let $g\colon \mathbb{R} \to [0,\infty)$ be measurable with $\int_{\mathbb{R}} g(y)dy \le 1$. Then by Fubini's Theorem and Hölder's inequality for y, Theorem 1.99 (for Lebesgue measure)

$$\int_{\mathbb{R}} g(y) \Big(\int_{\mathbb{R}} f(x,y)dx \Big) dy = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} f(x,y)g(y)dy \Big) dx \le \int_{\mathbb{R}} \|f(x,y)\|_{\infty,dy} \, dx. \qquad (*)$$

From the Reverse Hölder inequality, if $h\colon \mathbb{R} \to \mathbb{R}$ is measurable, then

$$\|h\|_\infty = \sup_{\substack{g\colon \mathbb{R}\to[0,\infty) \\ \int_{\mathbb{R}} g(y)dy \le 1}} \int_{\mathbb{R}} g(x)h(x)dx.$$

So, taking the supremum over such g in $(*)$, $\big\| \int_{\mathbb{R}} f(x,y)dx \big\|_{\infty,dy} \le \int_{\mathbb{R}} \|f(x,y)\|_{\infty,dy} \, dx$. □

We say $f\colon \mathbb{R} \to \mathbb{R}$ is a **Schwartz function** if, for any integers $j, k \ge 1$, f is k times continuously differentiable and there exists $c_{j,k} \in \mathbb{R}$ such that

$$\left| f^{(k)}(x) \right| \le \frac{c_{jk}}{1 + |x|^j}, \qquad \forall\, x \in \mathbb{R}.$$

Proposition 9.2 (Properties of Convolution on $\mathbb{R}$). *Let $1 \le p \le \infty$, let p' with $1/p + 1/p' = 1$. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ with $\int_{\mathbb{R}} |\phi(x)| \, dx < \infty$, let $\varepsilon > 0$ and define $\phi_\varepsilon(x) := \frac{1}{\varepsilon}\phi(x/\varepsilon)$ for any $x \in \mathbb{R}$ and $c := \int_{\mathbb{R}} \phi(x)dx$. Let $f, g\colon \mathbb{R} \to \mathbb{R}$ be Schwartz functions.*

(a) *For any $1 \le p < \infty$, $\lim_{\varepsilon \downarrow 0} \|\phi_\varepsilon * f - cf\|_p = 0$.*
(b) $\lim_{\varepsilon \to 0^+} \|\phi_\varepsilon * f - cf\|_\infty = 0$.
(c) *For any $x \in \mathbb{R}$, $\lim_{\varepsilon \to 0^+} (\phi_\varepsilon * f)(x) = cf(x)$ (using only that f is bounded, continuous).*
(d) *The convergence in (c) is uniform on $\mathbb{R}$ (using only that f is uniformly continuous).*
(e) *$\forall\, m \ge 1$, $f * g$ is m times continuously differentiable, and $(f * g)^{(m)} = f^{(m)} * g$.*

Proof of (a),(b):

$$\begin{aligned}\|\phi_\varepsilon * f - cf\|_p &= \left\| \int_{\mathbb{R}} \phi_\varepsilon(y)(f(x-y)-f(x))dy \right\|_{p,dx} \\ &\le \int_{\mathbb{R}} |\phi_\varepsilon(y)| \, \|f(x-y)-f(x)\|_{p,dx} \, dy \quad \text{, by Theorem. 9.1} \\ &= \int_{\mathbb{R}} |\phi(y)| \, \|f(x-\varepsilon y)-f(x)\|_{p,dx} \, dy, \text{ changing variables.}\end{aligned}$$

The y-integrand is bounded by $2\|f\|_p \int_{\mathbb{R}} |\phi(y)|\, dy < \infty$ and by $|\phi(y)|\,|\varepsilon y|\,\|f'\|_\infty$ by the Fundamental Theorem of Calculus. Since f is Schwartz, the latter quantity is bounded, so it goes to zero pointwise as $\varepsilon \to 0$. So, the Dominated Convergence Theorem, Theorem 3.10, implies (a) and (b).

Proof of (c): Arguing as in (a) (taking absolute values, changing variables, and applying Dominated Convergence),

$$|(\phi_\varepsilon * f)(x) - cf(x)| \le \int_{\mathbb{R}} |\phi(y)|\,|f(x-\varepsilon y)-f(x)|\,dy \to 0.$$

Proof of (d): Let $\eta > 0$. Choose $m > 0$ so that $2\|f\|_\infty \int_{|y|>m} |\phi(y)| \le \eta$. Choose $\delta > 0$ by uniform continuity of f so that for any $x \in \mathbb{R}$, if $|u| \le \delta$ then $|f(x+u)-f(x)| \le \eta/\|\phi\|_1$. Then for any $0 < \varepsilon \le \delta/m$ and for any $x \in \mathbb{R}$, if $|y| \le m$, then $|f(x-\varepsilon y)-f(x)| \le \eta/\|\phi\|_1$. So, continuing the calculation of (c), and applying the definition of m,

$$\begin{aligned}\int_{\mathbb{R}} |\phi(y)|\,|f(x-\varepsilon y)-f(x)|\,dy &= \int_{\{y\in\mathbb{R}:\,|y|>m\}} (\cdots) + \int_{\{y\in\mathbb{R}:\,|y|\le m\}} (\cdots) \\ &\le 2\|f\|_\infty \int_{\{y\in\mathbb{R}:\,|y|>m\}} |\phi(y)|\,dy + \int_{\{y\in\mathbb{R}:\,|y|\le m\}} |\phi(y)|\,\frac{\eta}{\|\phi\|_1} \le \eta+\eta = 2\eta.\end{aligned}$$

Proof of (e): Let $h > 0$ and $x \in \mathbb{R}$. Then

$$\begin{aligned}\left|\frac{(f*g)(x+h)-(f*g)(x)}{h} - (f'*g)(x)\right| &\le \left\|\frac{f(x+h)-f(x)}{h} - f'(x)\right\|_{\infty,dx} \|g\|_1 \\ &\le \left\|\frac{1}{h}\int_x^{x+h} (x+h-t)f''(t)dt\right\|_{\infty,dx} \|g\|_1 \le |h|\,\|f''\|_\infty \|g\|_1.\end{aligned}$$

Since f is a Schwartz function, $\|f''\|_\infty < \infty$, so the case $m = 1$ follows by letting $h \to 0^+$. The case of larger m follows by iteration. □

Let $f\colon \mathbb{R} \to \mathbb{R}$ with $\int_{\mathbb{R}} |f(x)|\,dx < \infty$. For any $\xi \in \mathbb{R}$, we define

$$\widehat{f}(\xi) = \mathcal{F}(f)(\xi) := \int_{\mathbb{R}} e^{ix\xi} f(x)dx.$$

Then $\widehat{f}\colon \mathbb{R} \to \mathbb{R}$ is called the **Fourier Transform** of f.

Proposition 9.3 (Properties of Fourier Transform). *Let f, g be Schwartz functions. Let $\xi \in \mathbb{R}$ and let $\lambda > 0$.*

(a) $|\widehat{f}(\xi)| \le \int_{\mathbb{R}} |f(x)|\,dx$, $\forall\ \xi \in \mathbb{R}$.

(b) $\mathcal{F}[f(x-h)](\xi) = e^{i\xi h}\widehat{f}(\xi)$, $\mathcal{F}[e^{ixh}f(x)](\xi) = \widehat{f}(\xi+h)$, $\forall\ h \in \mathbb{R}$.
(c) $\mathcal{F}[f(x/\lambda)](\xi) = \lambda\widehat{f}(\lambda\xi)$.
(d) $\widehat{(f*g)} = \widehat{f}\,\widehat{g}$
(e) $\partial\widehat{f}/\partial\xi = \mathcal{F}(ixf(x))$
(f) $\mathcal{F}[f'](\xi) = -i\xi\widehat{f}(\xi)$.
(g) $\int_{\mathbb{R}} f(x)\widehat{g}(x)dx = \int_{\mathbb{R}} \widehat{f}(x)g(x)dx$.

Proof of (a): $|\widehat{f}(\xi)| = \left|\int_{\mathbb{R}} e^{ix\xi}f(x)dx\right| \le \int_{\mathbb{R}} |f(x)|\,dx$.
Proof of (b): By the change of variables formula, if $\xi \in \mathbb{R}$,

$$\mathcal{F}[f(x-h)](\xi) = \int_{\mathbb{R}} e^{ix\xi}f(x-h)dx = e^{ixh}\int_{\mathbb{R}} e^{ix\xi}f(x)dx = e^{ixh}\widehat{f}(\xi).$$

$$\mathcal{F}[e^{ixh}f(x)](\xi) = \int_{\mathbb{R}} e^{ix(\xi+h)}f(x)dx = \widehat{f}(\xi+h).$$

Proof of (c): By the change of variables formula,

$$\mathcal{F}[f(x/\lambda)](\xi) = \int_{\mathbb{R}} e^{ix\xi}f(x/\lambda)dx = \lambda\int_{\mathbb{R}} e^{ix\xi\lambda}f(x)dx = \lambda\widehat{f}(\xi\lambda).$$

Proof of (d): Applying Fubini's Theorem, Theorem 1.79, and part (b) give

$$\int_{\mathbb{R}} e^{ix\xi}\left(\int_{\mathbb{R}} f(x-y)g(y)dy\right)dx = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{ix\xi}f(x-y)dxg(y)dy$$
$$\overset{(b)}{=} \int_{\mathbb{R}} e^{i\xi y}\widehat{f}(\xi)g(y)dy = \widehat{f}(\xi)\int_{\mathbb{R}} e^{i\xi y}g(y)dy = \widehat{f}(\xi)\widehat{g}(\xi).$$

Proof of (e): Let $h > 0$. Using part (b) and the Dominated Convergence Theorem 3.10,

$$\frac{\widehat{f}(\xi+h)-\widehat{f}(\xi)}{h} \overset{(b)}{=} \mathcal{F}\left[\left(\frac{e^{ixh}-1}{h}\right)f(x)\right](\xi) \to \mathcal{F}[ixf(x)](\xi)\,, \text{as } h \to 0.$$

We now justify the use of the Dominated Convergence Theorem. By the Mean Value Theorem, $\left|\mathrm{Re}(e^{ixh}-1)/h\right| = |(\cos(xh)-1)/h| \le |x|$ and $\left|\mathrm{Im}(e^{ixh}-1)/h\right| = |(\sin(xh)-1)/h| \le |x|$, so $\left|(e^{ixh}-1)/h\right| \le 2\,|x|$ and $\left|f(x)(e^{ixh}-1)/h\right| \le 2\,|x|\,|f(x)|$.
Proof of (f): Integrating by parts and then using that f is a Schwartz function

$$\mathcal{F}[f'(x)](\xi) = \lim_{N\to\infty}\int_{-N}^{N} f'(x)e^{ix\xi}dx = \lim_{N\to\infty} -\int_{-N}^{N} f(x)(i\xi)e^{ix\xi}dx = -i\xi\widehat{f}(\xi).$$

Proof of (g): Apply Fubini's Theorem 1.79. □

Proposition 9.4. *Let f, g be Schwartz functions. Let $\xi \in \mathbb{R}$.*
(a) $\mathcal{F}[e^{-x^2/2}](\xi) = \sqrt{2\pi}e^{-\xi^2/2}$.
(b) $\lim_{\xi\to\infty}\widehat{f}(\xi) = 0$.
(c) *$\widehat{f}$ is a Schwarz function.*

Proof. Let $\xi \in \mathbb{R}$. Completing the square, and then shifting the contour in the complex plane,

$$\int_{\mathbb{R}} e^{-x^2/2+ix\xi}dx = e^{-\xi^2/2}\int_{\mathbb{R}} e^{-(x-i\xi)^2/2}dx = e^{-\xi^2/2}\int_{\mathbb{R}} e^{-x^2/2}dx = \sqrt{2\pi}e^{-\xi^2/2}.$$

Now, let $\phi(x) := e^{-x^2/2}/\sqrt{2\pi}$ for any $x \in \mathbb{R}$ and denote $\phi_\varepsilon(x) := \varepsilon^{-1}\phi(x/\varepsilon)$ for any $x \in \mathbb{R}$. Note that $\int_{\mathbb{R}} \phi_\varepsilon(x)dx = 1$. From Proposition 9.3(a),(d) and Proposition 9.2(a),

$$\left|\widehat{\phi_\varepsilon}(\xi)\widehat{f}(\xi) - \widehat{f}(\xi)\right| = \left|\widehat{\phi_\varepsilon * f}(\xi) - \widehat{f}(\xi)\right| \leq \int_{\mathbb{R}} |\phi_\varepsilon * f(x) - f(x)|\, dx \to 0,$$

as $\varepsilon \to 0$. Combining this statement with Proposition 9.3(c) and part (a) of the current Proposition, $e^{-\varepsilon^2\xi^2/2}\widehat{f}(\xi)$ converges to $\widehat{f}(\xi)$ uniformly over all $\xi \in \mathbb{R}$, as $\varepsilon \to 0$. Since $\widehat{f}$ itself is bounded by Proposition 9.3(a), $e^{-\varepsilon^2\xi^2/2}\widehat{f}(\xi)$ vanishes at $\xi = \infty$, for every $\varepsilon > 0$. So, the uniform convergence implies that $\widehat{f}(\xi)$ also vanishes as $\xi \to \infty$, proving (b).

To prove (c), note that repeated application of Proposition 9.3 shows that $\widehat{f}$ is k times differentiable for any $k \geq 1$, since f is a Schwartz function. And part (b) of the current Proposition says that $f^{(k)}$ vanishes at infinity for any $k \geq 1$, so repeated application of Proposition 9.3(f) shows that $\widehat{f}$ is a Schwartz function. □

Exercise 9.5. Give an alternate proof of the fact $\mathcal{F}[e^{-x^2/2}](\xi) = \sqrt{2\pi}e^{-\xi^2/2}$ using the following strategy:

- Let $g(\xi) := (2\pi)^{-1/2}\mathcal{F}[e^{-x^2/2}](\xi)$. Show that $g'(\xi) = -\xi g(\xi)$ for all $\xi \in \mathbb{R}$.
- Deduce that $(d/d\xi)(g(\xi)e^{\xi^2/2}) = 0$.
- Finally, conclude that $g(\xi) = e^{-\xi^2/2}$.

Theorem 9.6 (Fourier Inversion). *Let $f\colon \mathbb{R} \to \mathbb{R}$ be a Schwartz function. Then*

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-ix\xi}\widehat{f}(\xi)d\xi, \qquad \forall\, x \in \mathbb{R}.$$

Proof. let $\phi(x) := e^{-x^2/2}/\sqrt{2\pi}$ for any $x \in \mathbb{R}$ and denote $\phi_\varepsilon(x) := \varepsilon^{-1}\phi(x/\varepsilon)$ for any $x \in \mathbb{R}$. Note that $\int_{\mathbb{R}} \phi_\varepsilon(x)dx = 1$. By Proposition 9.3(c) and Proposition 9.4(a), $\mathcal{F}[\phi](\xi) = e^{-\xi^2/2}$, $\mathcal{F}[\phi_\varepsilon](\xi) = e^{-\varepsilon^2\xi^2/2}$, and $\mathcal{F}(\mathcal{F}(\phi_\varepsilon)) = 2\pi\phi_\varepsilon$. So, using Theorem 9.3(g), we get

$$2\pi\int_{\mathbb{R}} f(x)\phi_\varepsilon(x)dx = \int_{\mathbb{R}} \widehat{f}(\xi)e^{-\varepsilon^2\xi^2/2}d\xi. \qquad (*)$$

Using this equality for $f(x+y)$, applying Theorem 9.3(b), and using $\phi_\varepsilon(-y) = \phi_\varepsilon(y)\ \forall\, y \in \mathbb{R}$,

$$\frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(\xi)e^{-ix\xi}e^{-\varepsilon^2\xi^2/2}d\xi \overset{(*)}{=} \int_{\mathbb{R}} f(x+y)\phi_\varepsilon(y)dy = \int_{\mathbb{R}} f(x-y)\phi_\varepsilon(y)dy = (\phi_\varepsilon * f)(x).$$

As $\varepsilon \to 0$, the left side converges to $\frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(\xi)e^{ix\xi}d\xi$ by the Dominated Convergence Theorem 3.10. And the right side tends to f uniformly in x by Proposition 9.2(d). So $f(x) = \frac{1}{2\pi}\int \widehat{f}(\xi)e^{-ix\xi}d\xi$ almost everywhere in $x \in \mathbb{R}$, hence everywhere since f is Schwartz. □

Lemma 9.7 (Stirling's Formula). *Let $n \in \mathbb{N}$. Then $n! \sim \sqrt{2\pi n}n^n e^{-n}$. That is,*

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}n^n e^{-n}} = 1.$$

Proof. We prove the weaker estimate that $\exists\ c \in \mathbb{R}$ such that

$$n! = (1 + O(1/n))e^{1-c}\sqrt{n}n^n e^{-n}. \qquad (*)$$

Note that $\log(n!) = \sum_{m=1}^{n} \log m$. We use integral comparison for this sum. On the interval $[m, m+1]$ the function $x \mapsto \log x$ has second derivative $O(1/m^2)$. So, Taylor expansion (i.e. the trapezoid rule) gives

$$\int_m^{m+1} \log x dx = \frac{1}{2}\log(m+1) + \frac{1}{2}\log m + O(1/m^2).$$

$$\int_1^n \log x dx = \sum_{m=1}^{n-1} \int_m^{m+1} \log x dx = \sum_{m=1}^{n-1} \log m + \frac{1}{2}\log n + c + O(1/n).$$

Since $\int_1^n \log x dx = n(\log(n) - 1) + 1$, $\log(n!) = \sum_{m=1}^{n} \log m$, exponentiating proves $(*)$. $\square$

Proposition 9.8 (**Differentiating under the Integral Sign**). *Let $f\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$. Suppose*

- *For all $\theta \in \mathbb{R}$, $\int_{\mathbb{R}^n} |f(\theta, x)|\, dx < \infty$.*
- *For almost all $\theta \in \mathbb{R}$, the derivative $\partial f(\theta, x)/\partial\theta$ exists for all $x \in \mathbb{R}^n$.*
- *There is a function $g\colon \mathbb{R}^n \to [0, \infty)$ with $\int_{\mathbb{R}^n} |g(x)|\, dx < \infty$ and $|\partial f(\theta, x)/\partial\theta| \leq g(x)$ for all $\theta \in \mathbb{R}$, $x \in \mathbb{R}^n$.*

Then for all $\theta \in \mathbb{R}$,

$$\frac{\partial}{\partial\theta} \int_{\mathbb{R}^n} f(\theta, x) dx = \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta} f(\theta, x) dx.$$

Proof. Let $h(\theta, x) := \frac{\partial}{\partial\theta} f(\theta, x)$ and let $h_0(\theta, x) := \int_0^\theta h(t, x) dt$ for any $\theta \in \mathbb{R}$, $x \in \mathbb{R}^n$. By assumption, $\int_{\mathbb{R}^n} |h(\theta, x)|\, dx < \infty$ for any $\theta \in \mathbb{R}$, so that $\int_0^\theta \int_{\mathbb{R}^n} |h(t, x)|\, dx dt < \infty$ for any $\theta \in \mathbb{R}$. By Fubini's Theorem 1.79,

$$\int_0^\theta \int_{\mathbb{R}^n} h(t, x) dx dt = \int_{\mathbb{R}^n} \int_0^\theta h(t, x) dt dx = \int_{\mathbb{R}^n} h_0(\theta, x) dx < \infty.$$

Taking derivatives in θ of both sides and applying Lebesgue's Fundamental Theorem of Calculus, Theorem 1.42 (twice) concludes the proof. $\square$

10. Appendix: Convergence in Distribution, Characteristic Functions

Definition 10.1 (**Vague Convergence of Measures**). Let $\mu, \mu_1, \mu_2, \ldots$ be a sequence of finite measures on $\mathbb{R}$ (i.e. $\mu(\mathbb{R}), \mu_n(\mathbb{R}) < \infty$ for all $n \geq 1$). We say that $\mu_1, \mu_2, \ldots$ **converges vaguely** (or **converges weakly**, or **converges in the weak* topology**) to μ if, for any continuous compactly supported function $g\colon \mathbb{R} \to \mathbb{R}$,

$$\lim_{n\to\infty} \int_{\mathbb{R}} g(x) d\mu_n(x) = \int_{\mathbb{R}} g(x) d\mu(x).$$

In functional analysis, there is a subtle but important distinction between weak and weak* convergence, though this difference of terminology seems to be ignored in the probability literature.

As we will show below, convergence in distribution of random variables $X_1, X_2, \ldots$ to a random variable X is equivalent to $\mu_{X_1}, \mu_{X_2}, \ldots$ converging vaguely to μ_X.

Proposition 10.2. *Let $X, X_1, X_2, \ldots$ be random variables with values in $\mathbb{R}$. Then the following are equivalent*

- *$X_1, X_2, \ldots$ converges in distribution to X.*

- $\mu_{X_1}, \mu_{X_2}, \ldots$ *converges vaguely to* μ_X.

Proof. Assume that $X_1, X_2, \ldots$ converges in distribution to X. Let $g\colon \mathbb{R} \to \mathbb{R}$ be a continuous compactly supported function. Then g is uniformly continuous. So, if $\varepsilon > 0$, there exist $t_1 < \cdots < t_m$ and $c_1, \ldots, c_m \in \mathbb{R}$ such that $g_\varepsilon(t) := \sum_{i=1}^{m-1} c_i 1_{(t_i, t_{i+1}]}(t)$ satisfies $|g_\varepsilon(t) - g(t)| < \varepsilon$ for all $t \in \mathbb{R}$. Since $F_X\colon \mathbb{R} \to [0,1]$ is monotone increasing and bounded, any point of discontinuity of F_X is a jump discontinuity. So, F_X has at most a countable set of points of discontinuity. Therefore, $t_1 < \cdots < t_m$ can be chosen to all be points of continuity of F_X. By the definition of the expected value,

$$\left| \mathbf{E}g(X) - \sum_{i=1}^{m-1} c_i \Big(F_X(t_{i+1}) - F_X(t_i) \Big) \right| = |\mathbf{E}g(X) - \mathbf{E}g_\varepsilon(X)| \le \mathbf{E}\,|g(X) - g_\varepsilon(X)| \le \varepsilon.$$

The same holds replacing X with any of $X_1, X_2, \ldots$. So, applying the triangle inequality,

$$\begin{aligned} &\limsup_{n\to\infty} |\mathbf{E}g(X_n) - \mathbf{E}g(X)| \\ &\quad \le \limsup_{n\to\infty} |\mathbf{E}g(X_n) - \mathbf{E}g_\varepsilon(X_n)| + |\mathbf{E}g_\varepsilon(X_n) - \mathbf{E}g_\varepsilon(X)| + |\mathbf{E}g_\varepsilon(X) - \mathbf{E}g(X)| \\ &\quad \le 2\varepsilon + \limsup_{n\to\infty} \sum_{i=1}^{m-1} |c_i|\,|F_{X_n}(t_{i+1}) - F_X(t_{i+1}) - [F_{X_n}(t_i) - F_X(t_i)]| = 2\varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$ as desired.

Now, suppose for any continuous, compactly supported $g\colon \mathbb{R} \to \mathbb{R}$, $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$. Let $t \in \mathbb{R}$ be a point of continuity of F_X. Then, for any $\varepsilon > 0$, there exists $\delta > 0$ such that if $|s - t| < 2\delta$, then $|F_X(s) - F_X(t)| < \varepsilon$. By continuity of the probability law, let $m > 0$ such that $\mathbf{P}(|X| > m) < \varepsilon$. By choice of δ, ε we have $\mathbf{P}(|X - t| < \delta) < \varepsilon$. Let $g\colon \mathbb{R} \to [0,1]$ so that $g = 0$ on $(-\infty, -2m]$, $g = 1$ on $(-m, t - \delta]$, $g = 0$ on (t, ∞) and g is linear otherwise. Then

$$\begin{aligned} \mathbf{E}g(X) &= \mathbf{E}g(X)(1_{-2m<X\le -m} + 1_{-m<X\le t-\delta} + 1_{t-\delta<X\le t}) \\ &= O(\varepsilon) + F_X(t - \delta) + O(\varepsilon) = F_X(t) + O(\varepsilon). \end{aligned}$$

Since $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$, there exists $n_0 = n_0(\varepsilon) > 0$ such that, for all $n > n_0$, $\mathbf{E}g(X_n) = F_X(t) + O(\varepsilon)$. By the definition of g,

$$\mathbf{P}(X_n \le t) \ge \mathbf{E}g(X_n) \ge F_X(t) - O(\varepsilon), \qquad \forall\, n > n_0(\varepsilon).$$

Repeating the above with g where $g = 1$ on $(t + \delta, m]$ and $g = 0$ on $(-\infty, t] \cup [2m, \infty)$ gives

$$\mathbf{P}(X_n > t) \ge 1 - F_X(t) - O(\varepsilon), \qquad \forall\, n > n_0(\varepsilon).$$

Combining these inequalities gives

$$F_{X_n}(t) = F_X(t) + O(\varepsilon), \qquad \forall\, n > n_0(\varepsilon).$$

Letting $\varepsilon \to 0^+$ concludes the proof. □

Lemma 10.3. *Let $\mu_1, \mu_2, \ldots$ be a sequence of probability measures on $\mathbb{R}$. Then any subsequential limit of the sequence (with respect to vague convergence) is a probability measure if and only if $\mu_1, \mu_2, \ldots$ is* ***tight****: $\forall\ \varepsilon > 0$, $\exists\ m = m(\varepsilon) > 0$ such that*

$$\limsup_{n\to\infty} (1 - \mu_n([-m, m])) \le \varepsilon.$$

Exercise 10.4. Let $X, X_1, X_2, \dots$ and let $Y, Y_1, Y_2, \dots$ be random variables with values in $\mathbb{R}$.

(i) Assume that X is constant almost surely. Show that $X_1, X_2, \dots$ converges to X in distribution if and only if $X_1, X_2, \dots$ converges to X in probability.
(ii) Prove Lemma 10.3.
(iii) Suppose that $X_1, X_2, \dots$ converges in distribution to X. Show there exist random variables $Z, Z_1, Z_2, \dots : \Omega \to \mathbb{R}$ such that $\mu_Z = \mu_X$, $\mu_{Z_n} = \mu_{X_n}$ for any $n \geq 1$, and such that $Z_1, Z_2, \dots$ converges almost surely to Z. (Hint: use Exercise 4.20.)
(iv) (Slutsky's Theorem) Suppose $X_1, X_2, \dots$ converges in distribution to X and $Y_1, Y_2, \dots$ converges in probability to Y. Assume Y is constant almost surely. Show that $X_1+Y_1, X_2+Y_2, \dots$ converges in distribution to $X+Y$. Show also that $X_1Y_1, X_2Y_2, \dots$ converges in distribution to XY. (Hint: either use (iii) or use (ii) to control error terms.) What happens if Y is not constant almost surely?
(v) (Fatou's lemma) If $g\colon \mathbb{R} \to [0, \infty)$ is continuous, and if $X_1, X_2, \dots$ converges in distribution to X, show that $\liminf_{n\to\infty} \mathbf{E}g(X_n) \geq \mathbf{E}g(X)$.
(vi) (Bounded convergence) If $g\colon \mathbb{R} \to \mathbb{C}$ is continuous and bounded, and if $X_1, X_2, \dots$ converges in distribution to X, show that $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$.
(vii) (Dominated convergence) If $X_1, X_2, \dots : \Omega \to \mathbb{R}$ converges in distribution to X, and if there exists a random variable $Y\colon \Omega \to [0, \infty)$ with $|X_n| \leq Y$ for all $n \geq 1$ and $\mathbf{E}Y < \infty$, show that $\lim_{n\to\infty} \mathbf{E}X_n = \mathbf{E}X$.

Theorem 10.5 (Lévy Continuity Theorem, Special Case). *Let $X, X_1, X_2, \dots$ be real-valued random variables (possibly on different sample spaces). The following are equivalent.*

- *For every $t \in \mathbb{R}$, $\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$.*
- *$X_1, X_2, \dots$ converges in distribution to X.*

Proof. The second condition implies the first by Exercise 10.4(vi).

Now, assume the first condition holds. Let $g\colon \mathbb{R} \to \mathbb{R}$ be a Schwartz function (for any integers $j, k \geq 1$, g is k times continuously differentiable and there exists $c_{j,k} \in \mathbb{R}$ such that $|g^{(k)}(x)| \leq \frac{c_{jk}}{1+|x|^j}$, $\forall\, x \in \mathbb{R}$.) The Fourier Inversion Formula, Theorem 9.6, implies that

$$g(X_n) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-iX_n y} \widehat{g}(y) dy.$$

where $\widehat{g}(y) = \int_{\mathbb{R}} e^{ixy} g(x) dx$ for all $y \in \mathbb{R}$. From the Fubini Theorem 1.79,

$$\mathbf{E}g(X_n) = \frac{1}{2\pi} \int_{\mathbb{R}} \mathbf{E}e^{-iX_n y} \widehat{g}(y) dy = \frac{1}{2\pi} \int_{\mathbb{R}} \phi_{X_n}(-y) \widehat{g}(y) dy.$$

Similarly, $\mathbf{E}g(X) = \frac{1}{2\pi} \int_{\mathbb{R}} \phi_X(-y) \widehat{g}(y) dy$. So, $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$ by the Dominated Convergence Theorem, Theorem 3.10 (and Proposition 9.4(c)). Since any continuous, compactly supported function g can be uniformly approximated by Schwartz functions in the L_∞ norm (by e.g. replacing g with $g * \phi_\varepsilon$, where $\phi_\varepsilon(x) = \varepsilon^{-1} e^{-x^2/(2\varepsilon^2)}/\sqrt{2\pi}$, letting $\varepsilon \to 0^+$ and applying Proposition 9.2(d)), the identity $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$ holds for any continuous, compactly supported $g\colon \mathbb{R} \to \mathbb{R}$. We then conclude by Proposition 10.2. □

Remark 10.6. In particular, if $Y = X_1 = X_2 = \cdots$, the above Theorem implies that if $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$, then X and Y have the same distribution.

Exercise 10.7 (**Lévy Continuity Theorem**). Let $X, X_1, X_2, \ldots$ be real-valued random variables (possibly on different sample spaces). Assume that, $\forall\, t \in \mathbb{R}$, $\phi(t) := \lim_{n\to\infty} \phi_{X_n}(t)$ exists. Then the following are equivalent.

(i) ϕ is continuous at 0.
(ii) $\mu_{X_1}, \mu_{X_2}, \ldots$ is tight. ($\forall\ \varepsilon > 0$, $\exists\ m = m(\varepsilon) > 0$ such that $\limsup_{n\to\infty}(1 - \mu_{X_n}([-m, m])) \le \varepsilon$.)
(iii) There exists a random variable X such that $\phi_X = \phi$.
(iv) $X_1, X_2, \ldots$ converges in distribution to X.

(Hint: Use Lemma 10.3 to get from (ii) to other conditions.)

11. Appendix: Moment Generating Functions

Exercise 11.1. Unfortunately, there exist random variables X, Y such that $\mathbf{E}X^n = \mathbf{E}Y^n$ for all $n = 1, 2, 3, \ldots$, but such that X, Y do not have the same CDF. First, explain why this does not contradict the Lévy Continuity Theorem, Weak Form. Now, let $-1 < a < 1$, and define a density

$$f_a(x) := \begin{cases} \frac{1}{x\sqrt{2\pi}} e^{-\frac{(\log x)^2}{2}} (1 + a\sin(2\pi \log x)) & \text{, if } x > 0 \\ 0 & \text{, otherwise.} \end{cases}$$

Suppose X_a has density f_a. If $-1 < a, b < 1$, show that $\mathbf{E}X_a^n = \mathbf{E}X_b^n$ for all $n = 1, 2, 3, \ldots$. (Hint: write out the integrals, and make a change of variables $s = \log(x) - n$.)

Theorem 11.2 (**Inversion of Moment Generating Function**). *Let X, Y be random variables. Denote $M_X(t) := \mathbf{E}e^{tX}$ for any $t \in \mathbb{R}$. Suppose $M_X(t) = M_Y(t)$ for all $t \in (-\varepsilon, \varepsilon)$. Then X and Y have the same distribution.*

Proof. From (the proof of) Lemma 3.8 with $\mu = \mathbf{P}$, $h = 1$, $k = 1$, $t(x) = x$, $M_X(t)$ is complex-differentiable in a neighborhood of the origin. From a well-known theorem from complex analysis, $M_X(z)$ is then equal to its power series for all $z \in \mathbb{C}$ with $|z| < \varepsilon$. That is, its power series is absolutely convergence for all $|z| < \varepsilon$, and

$$M_X(z) = \sum_{k=0}^{\infty} \frac{(d/dt)^k|_{t=0} M_X(t)}{k!} z^k, \qquad \forall\ |z| < \varepsilon.$$

By Lemma 3.8 again, $(d/dt)^k|_{t=0} M_X(t) = \mathbf{E}X^k$ for all $k \ge 0$. Since the series converges absolutely, we have

$$\lim_{k\to\infty} \frac{\mathbf{E}X^k}{k!} x^k = 0, \qquad \forall\, 0 < x < \varepsilon. \qquad (*)$$

Fix $0 < r < s < \varepsilon$. If k is an odd integer, then $(k+1)r^k < \varepsilon^{k+1}$ for sufficiently large k, and for all $0 < x < r$, $|x|^k \le 1 + |x|^{k+1}$, so multiplying these inequalities and taking expected values gives

$$\frac{\mathbf{E}\,|X|^k\, r^k}{k!} \le \frac{r^k}{k!} + \frac{\mathbf{E}\,|X|^{k+1}\, s^{k+1}}{(k+1)!}.$$

That is, $(*)$ implies that

$$\lim_{k\to\infty} \frac{\mathbf{E}\,|X|^k}{k!} x^k = 0, \qquad \forall\, 0 < x < \varepsilon. \qquad (**)$$

Let $i := \sqrt{-1}$. Let $x, t, h \in \mathbb{R}$. From the Taylor expansion of the exponential function,

$$\left| e^{itx} \left(e^{ihx} - \sum_{k=0}^{n} \frac{(ihx)^n}{n!} \right) \right| = \left| e^{ihx} - \sum_{k=0}^{n} \frac{(ihx)^k}{k!} \right| \leq \frac{|hx|^{n+1}}{(n+1)!}.$$

We denote $\phi_X(t) := \mathbf{E}e^{itX}$. So, taking expected values of these same quantities with $x = X$,

$$\left| \phi_X(t+h) - \sum_{k=0}^{n} \frac{(i)^k \mathbf{E} e^{itX} X^k}{k!} \right| \leq \frac{|h|^{n+1} \mathbf{E}\,|X|^{n+1}}{(n+1)!}, \qquad \forall\, t \in \mathbb{R},\, \forall\, h \in (-\varepsilon, \varepsilon).$$

By $(**)$, the series then converges, so that

$$\phi_X(t+h) = \sum_{k=0}^{\infty} \frac{i^k \mathbf{E} e^{itX} X^k}{k!} h^k, \qquad \forall\, t \in \mathbb{R},\, \forall\, h \in (-\varepsilon, \varepsilon).$$

By Lemma 3.8, differentiating ϕ_X can occur under the expected value, so that

$$\phi_X(t+h) = \sum_{k=0}^{\infty} \frac{\phi_X^{(k)}(t)}{k!} h^k, \qquad \forall\, t \in \mathbb{R},\, \forall\, h \in (-\varepsilon, \varepsilon). \qquad (***)$$

Similarly,

$$\phi_Y(t+h) = \sum_{k=0}^{\infty} \frac{\phi_Y^{(k)}(t)}{k!} h^k, \qquad \forall\, t \in \mathbb{R},\, \forall\, h \in (-\varepsilon, \varepsilon). \qquad (\ddagger)$$

Setting $t = 0$, using these equalities and our assumption, we see that for any $k \geq 0$,

$$\frac{d^k}{dt^k}|_{t=0} \phi_X(t) = i^k \mathbf{E} X^k = i^k \frac{d^k}{dt^k}|_{t=0} \mathbf{E} e^{tX} = i^k \frac{d^k}{dt^k}|_{t=0} \mathbf{E} e^{tY} = \frac{d^k}{dt^k}|_{t=0} \mathbf{E} e^{itY}.$$

Therefore, $\phi_X(t) = \phi_Y(t)$ for all $t \in (-\varepsilon, \varepsilon)$ by $(***)$ and $(\ddagger)$, since each coefficient of their power series also agrees. Consequently, $\phi_X(t) = \phi_Y(t)$ for all $t \in (-2\varepsilon, 2\varepsilon)$ by $(***)$ and $(\ddagger)$. Iterating this argument, $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$. We then conclude by Remark 10.6. $\square$

12. Appendix: Notation

Let n, m be a positive integers. Let A, B be sets contained in a universal set Ω.

$\mathbb{N} = \{1, 2, \ldots\}$ denotes the set of natural numbers

$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ denotes the set of integers

$\mathbb{Q} = \{a/b \colon a, b, \in \mathbb{Z}, b \neq 0\}$ denotes the set of rational numbers

$\mathbb{R}$ denotes the set of real numbers

$\mathbb{C} = \{a + b\sqrt{-1} \colon a, b \in \mathbb{R}\}$ denotes the set of complex numbers

$\in$ means "is an element of." For example, $2 \in \mathbb{R}$ is read as "2 is an element of $\mathbb{R}$."

$\forall$ means "for all"

$\exists$ means "there exists"

$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \colon x_i \in \mathbb{R} \, \forall \, 1 \leq i \leq n\}$

$f \colon A \to B$ means f is a function with domain A and range B. For example,

$f \colon \mathbb{R}^2 \to \mathbb{R}$ means that f is a function with domain $\mathbb{R}^2$ and range $\mathbb{R}$

$\emptyset$ denotes the empty set

$A \subseteq B$ means $\forall \, a \in A$, we have $a \in B$, so A is contained in B

$A \setminus B := \{a \in A \colon a \notin B\}$

$A^c := \Omega \setminus A$, the complement of A in Ω

$A \cap B$ denotes the intersection of A and B

$A \cup B$ denotes the union of A and B

$A \Delta B := (A \setminus B) \cup (B \setminus A)$

$\mathbf{P}$ denotes a probability law on Ω

Let $n \geq m \geq 0$ be integers. We define

$$\binom{n}{m} := \frac{n!}{(n-m)!m!} = \frac{n(n-1)\cdots(n-m+1)}{m(m-1)\cdots(2)(1)}.$$

Let $a_1, \ldots, a_n$ be real numbers. Let n be a positive integer.

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_{n-1} + a_n.$$

$$\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdots a_{n-1} \cdot a_n.$$

$\min(a_1, a_2)$ denotes the minimum of a_1 and a_2.

$\max(a_1, a_2)$ denotes the maximum of a_1 and a_2.

The min of a set of nonnegative real numbers is the smallest element of that set. We also define $\min(\emptyset) := \infty$.

Let $A \subseteq \mathbb{R}$.

$\sup A$ denotes the supremum of A, i.e. the least upper bound of A.
$\inf A$ denotes the infimum of A, i.e. the greatest lower bound of A.

Let $X\colon \Omega \to \mathbb{R}$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mu)$.

$\mathbf{E}(X)$ denotes the expected value of X
$\|X\|_p := (\mathbf{E}\,|X|^p)^{1/p}$, denotes the L_p-norm of X when $1 \leq p < \infty$
$\|X\|_\infty := \inf\{c > 0\colon \mathbf{P}(|X| \leq c) = 1\}$, denotes the L_∞-norm of X
$\operatorname{var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2$, the variance of X
$\sigma_X = \sqrt{\operatorname{var}(X)}$, the standard deviation of X

Let $A \subseteq \Omega$.

$\mathbf{E}(X|A) := \mathbf{E}(X1_A)/\mathbf{P}(A)$ denotes the expected value of X conditioned on the event A.

$1_A\colon \Omega \to \{0, 1\}$, denotes the indicator function of A, so that

$$1_A(\omega) = \begin{cases} 1 & \text{, if } \omega \in A \\ 0 & \text{, otherwise.} \end{cases}$$

Let X be a random variable on a sample space Ω, so that $X\colon \Omega \to \mathbb{R}$. Let $\mathbf{P}$ be a probability law on Ω. Let $x, t \in \mathbb{R}$.

$F_X(x) = \mathbf{P}(X \leq x) = \mathbf{P}(\{\omega \in \Omega\colon X(\omega) \leq x\})$
the Cumulative Distibution Function of X.
$M_X(t) = \mathbf{E}e^{tX}$ denotes the Moment Generating Function of X at $t \in \mathbb{R}$

Let $g, h\colon \mathbb{R} \to \mathbb{R}$. Let $t \in \mathbb{R}$.

$(g * h)(t) = \displaystyle\int_{-\infty}^{\infty} g(x)h(t-x)dx$ denotes the convolution of g and h at $t \in \mathbb{R}$

Let n, k be a positive integers and let μ be a measure on $\mathbb{R}^n$. Let $t_1, \ldots, t_k\colon \mathbb{R}^n \to \mathbb{R}$. Let $h\colon \mathbb{R}^n \to [0, \infty]$ so that h is not identically zero. Let $\Theta \subseteq \mathbb{R}^k$ and let $w\colon \Theta \to \mathbb{R}^k$. For any $\theta \in \Theta$ define

$$a(w(\theta)) := \log \int_{\mathbb{R}^n} h(x) \exp\Big(\sum_{i=1}^{k} w_i(\theta)t_i(x)\Big) d\mu(x).$$

We define a k**-parameter exponential family** to be a set of functions $\{f_\theta \colon \theta \in \Theta,\, a(w(\theta)) < \infty\}$, where

$$f_\theta(x) := h(x) \exp\Big(\sum_{i=1}^{k} w_i(\theta) t_i(x) - a(w(\theta)) \Big), \qquad \forall\, x \in \mathbb{R}^n.$$

Let $\theta \in \Theta$

$\mathbf{P}_\theta$	denotes probability law corresponding to f_θ.
$\mathbf{E}_\theta$	denotes expected value with respect to f_θ.

USC Mathematics, Los Angeles, CA
Email address: stevenmheilman@gmail.com