MATH 408, MATHEMATICAL STATISTICS, FALL 2023

STEVEN HEILMAN

Contents

Date: November 19, 2023

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 $-\infty \leq a \leq b \leq \infty$, 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 $-\infty \leq a \leq b \leq \infty$, we have

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

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

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

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 $m \geq 1$ 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

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

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 the sum 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

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

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 $-\infty \leq a < b \leq \infty$ and $-\infty \leq c < d \leq \infty$, 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). 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

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

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 $1 \leq i \leq n$. 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 $i \neq j$. 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 $\text{cov}(X, Y)$, by

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

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

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

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

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

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

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

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

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

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 $\mathbf{P}$,

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

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

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

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

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,

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

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. □

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 $y \in \mathbb{R}$,

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

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

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

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*

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

(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,

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

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

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

□

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

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

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

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 $0 < p \le \infty$. 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.

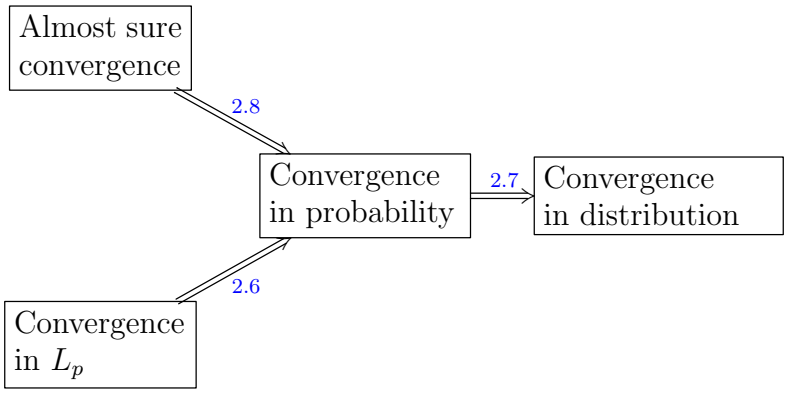
Then, show that the converse is false.

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 3.

Theorem 2.10 (**Weak Law of Large Numbers**). *Let $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ 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 $X_1, X_2, \dots : \Omega \to \mathbb{R}$ 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

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

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

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 $X : \mathbb{R} \to \mathbb{R}$. 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**). *For any $n \geq 1$, let $X_{n,1}, \ldots, X_{n,n} \colon \Omega_n \to \mathbb{R}$ 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*

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

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

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

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**). *Let $\sigma > 0$. Let $X_1, X_2, \ldots$ be i.i.d. real-valued random variables with mean zero, $\mathbf{E}X_1^2 = \sigma^2$, and $\mathbf{E}\,|X_1|^3 < \infty$. Let Z be a standard Gaussian random variable. Then for any $n \geq 1$,*

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

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 $X_1, X_2, \ldots$ 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 $X_1, X_2, \ldots$ 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 $X_1, X_2, \ldots$ 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 $X^{(1)}, X^{(2)}, \ldots$ 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 $X^{(n)} = (X_1^{(n)}, \ldots, X_d^{(n)}) \in \mathbb{R}^d$.) 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 $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ 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 $1 \leq i \leq d$.)

Exercise 2.36. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables that converge in probability to $X \colon \Omega \to \mathbb{R}$. Let $f \colon \mathbb{R} \to \mathbb{R}$ be continuous. Then $f(X_1), f(X_2), \ldots$ converges in probability to $f(X)$.

Proposition 2.37.

- *(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 2.38. 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 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 2.39. 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.)

Theorem 2.40 (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*

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

Exercise 2.41. Let $X : \Omega \to \mathbb{R}^n$ be a random variable with the **standard Gaussian distribution**:

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

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. 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 3.1 (**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 2.38, 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 3.2 (**Statistic**)**.** Let n, k be positive integers. Let $X_1, \ldots, X_n$ be a random sample of size n. Let $t\colon \mathbb{R}^n \to \mathbb{R}^k$. A **statistic** is a random variable of the form $Y := t(X_1, \ldots, X_n)$. The distribution of Y is called a **sampling distribution**.

Example 3.3. 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 3.4. 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 3.5. Let $n \geq 2$ be an integer. Let $X_1, \ldots, X_n$ be a random sample of size n. Assume that $\mu := \mathbf{E}X_1 \in \mathbb{R}$ and $\sigma := \sqrt{\operatorname{var}(X_1)} < \infty$. Let $\overline{X}$ be the sample mean and let S be the sample standard deviation of the random sample. Show the following

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

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 4.

Exercise 3.6. 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 $t = \mathbf{E}X$.

3.1. **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 3.7. *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 variables $\overline{X}$ is independent of all of the random variables $X_1 - \overline{X}, \ldots, X_n - \overline{X}$. This follows from Exercise 2.41, since the vector $(1, \ldots, 1) \in \mathbb{R}^n$ is orthogonal to any vector in the span of

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

(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 2.41 since the vectors $(1, 0, 0, \ldots) - \frac{1}{n}(1, \ldots, 1), (0, 1, 0, \ldots, 0) - \frac{1}{n}(1, \ldots, 1)$ are not orthogonal in $\mathbb{R}^n$.) So, the first item is be completed since S is a function of $X_1 - \overline{X}, \ldots, X_n - \overline{X}$.

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 3.8,

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

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. $\square$

Lemma 3.8. *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 3.9. *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 3.10. 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 $(\overline{X}-\mu)/(S/\sqrt{n})$ also has Student's t-distribution, since $\overline{X}-\mu$ has mean zero, 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 3.7).

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 $da=\frac{1}{2\sqrt{x}}dx$,

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

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

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

□

Exercise 3.11. 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 3.12 (Order Statistics). Let $X\colon \Omega \to \mathbb{R}$ 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, i.e. by differentiating $\mathbf{P}(X_{(j)} \le x) = \sum_{k=j}^{n} \binom{n}{k} F_X(x)^k (1-F_X(x))^{n-k}$, where $F_X(x) := \mathbf{P}(X \le x)$ for any $x \in \mathbb{R}$.)
- 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 $n-j+1$. 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).

3.2. **The Delta Method.** From Examples 3.3 and 3.4 and Exercise 3.5, 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 3.13 (**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, for all $s, t > 0$, $\lim_{n \to \infty} \mathbf{P}(|Y_n - \theta| > st/\sqrt{n}) = 2 \int_{st}^{\infty} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}}$. So, $\forall\ n \ge 1$,

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

As $n \to \infty$, the first term converges to $2 \int_{st}^{\infty} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}}$, and the second term goes to zero since $\lim_{z \to 0} (h(z)/z) = 0$. So, for any $s, t > 0$, $\lim_{n \to \infty} \mathbf{P}(\sqrt{n}\,|h(Y_n - \theta)| > t) \le 2 \int_{st}^{\infty} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}}$. 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 2.37 and $(*)$, $\sqrt{n}[f(Y_n) - f(\theta)]$ converges in distribution to a mean zero Gaussian with variance $\sigma^2 (f'(\theta))^2$. □

Example 3.14. Suppose $\overline{X}_n$ is the sample mean for a random sample $X_1, \ldots, X_n$ of size n and $0 < \operatorname{var}(X_1) < \infty$. Let $\mu := \mathbf{E}X_1 \ne 0$. From the Central Limit Theorem 2.13, $\sqrt{n}(\overline{X}_n - \mu)$ converges in distribution to a mean zero Gaussian with variance $\sigma^2 := \operatorname{var}(X_1)$.

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

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

by Theorem 3.15 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 3.15 (**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 3.16 (**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 $s, t > 0$, $\lim_{n\to\infty} \mathbf{P}(|Y_n - \theta| > st/\sqrt{n}) = 2\int_{st}^{\infty} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}}$. 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 $2\int_{st}^{\infty} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}}$, and the second term goes to zero since $\lim_{z\to 0}(h(z)/z^2) = 0$. So, for any $s, t > 0$, $\lim_{n\to\infty} \mathbf{P}(n\,|h(Y_n - \theta)| > t) \leq 2\int_{st}^{\infty} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}}$. 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 2.37 and $(*)$, $n[f(Y_n) - f(\theta)]$ converges in distribution to a chi squared random variable with one degree of freedom, multiplied by $\sigma^2 f''(\theta)$. □

Let $m > 2$ be an integer. Theorem 3.16 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 the absolute value of a Gaussian to the m^{th} power, multiplied by $\sigma^m \frac{1}{m!} f^{(m)}(\theta)$.

3.3. **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 3.17. 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 3.18 (**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

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

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

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

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

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

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 3.19. 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 3.19 then suggest the following method for simulating a random variable on a computer.

Algorithm 3.20 (**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 3.21. 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 3.19, F^{-1} is an exponential random variable with parameter 1 if $\mathbf{P}$ is the uniform probability law on $(0,1)$. Or by Algorithm 3.20, $F^{-1}(U) = -\log(1-U)$ is an exponential random variable with parameter 1.

When an explicit formula can be given for Y in Algorithm 3.20, the random variable can be simulated efficiently. However, if Y cannot be accurately or efficiently computed, Algorithm 3.20 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 3.20. 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 3.22 (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 $\{a_{ij}\}_{1 \leq i,j \leq n}$ be an $n \times n$ symmetric positive semidefinite matrix. That is, for any $v \in \mathbb{R}^n$, we have

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

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

- Let $X = (X_1, \ldots, X_n)$ 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
$$\mathbf{E}(Z_i Z_j) = a_{ij}, \qquad \forall\, 1 \le i, j \le n.$$

3.4. **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.

4. Estimation of Parameters

A basic problem in statistics is to fit data to an unknown probability distribution. As in Exercise 2.38, 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 general, we might want to estimate any number of unknown parameters.

Let $X_1, \dots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. We can regard $\{f_\theta \colon \theta \in \Theta\}$ as either a family of probability density functions, or a family of probability mass functions. 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**.

Example 4.1. In Exercise 2.38 we have a random sample $X_1, \dots, X_{20}$ from a Gaussian distribution with unknown mean and variance. We denote the unknown Gaussians as
$$\{f_\theta \colon \theta \in \Theta\} = \{f_{\mu,\sigma}(x) \colon (\mu, \sigma) \in \mathbb{R}^2,\, \mu \in \mathbb{R},\, \sigma > 0\} = \left\{ \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : \mu \in \mathbb{R},\, \sigma > 0 \right\}.$$
One estimator for the unknown mean μ is the sample mean
$$\frac{X_1 + \cdots + X_{20}}{20}.$$
A "less good" estimator for the unknown mean μ could be $X_1 + X_2$ or $(X_1 + X_3)/2$.

As previously discussed, an estimator for the unknown variance σ^2
$$\frac{1}{19} \sum_{i=1}^{20} (X_i - \overline{X})^2.$$
And an estimator for the unknown parameter σ itself is
$$S := \sqrt{\frac{1}{19} \sum_{i=1}^{20} (X_i - \overline{X})^2}.$$

As we see from this example, there are many ways of defining estimators for various unknown parameters. One focus of this course will be criteria for determining if an estimator is "good" or not.

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 4.2. 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 3.5 that the sample mean and sample variance are unbiased estimates of the mean and variance, respectively.

4.1. Method of Moments.

Definition 4.3 (Consistency). Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of distributions. Let $Y_1, Y_2, \dots$ be a sequence of estimators of $g(\theta)$ where $g \colon \Theta \to \mathbb{R}^k$. We say that $Y_1, Y_2, \dots$ is **consistent** for $g(\theta)$ if, for any $\theta \in \Theta$, $Y_1, Y_2, \dots$ 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 4.4. Let $X_1, \dots, X_n$ 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

$$M_j = M_j(\theta) := \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

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 $h \colon \mathbb{R}^k \to \mathbb{R}^k$ is continuous, and if $Y_1, Y_2, \dots$ is consistent for $g(\theta)$, then $h(Y_1), h(Y_2), \dots$ is consistent for $h(g(\theta))$ by Exercise 2.36.

Definition 4.5 (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

$$g(\theta) = h(\mu_1, \dots, \mu_j).$$

Then the estimator

$$h(M_1, \dots, M_j)$$

is a **method of moments** estimator for $g(\theta)$.

Example 4.6. 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 estimation 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). In the following example, we demonstrate an estimator with much smaller variance than the method of moments estimator.

Example 4.7. Suppose $X_1, \dots, X_n$ is a random sample of size n from the uniform distribution on the interval $[0, \theta]$ and $\theta > 0$ is unknown. 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 (by Example 4.4), but its variance is $\frac{1}{3n}\theta^2$. It turns out the estimator $(1+1/n)X_{(n)}$ is unbiased and consistent for θ with a smaller variance. From Definition 1.37 we have

$$\begin{aligned}
\mathbf{E}(1+1/n)X_{(n)} &= (1+1/n)\int_0^\theta \mathbf{P}(X_{(n)} > t)dt = (1+1/n)\int_0^\theta [1 - \mathbf{P}(X_{(n)} < t)]dt \\
&= (1+1/n)\int_0^\theta [1 - \mathbf{P}(X_{(n)} < t)]dt = (1+1/n)(\theta - \int_0^\theta \mathbf{P}(X_1 < t)^n dt) \\
&= (1+1/n)(\theta - \int_0^\theta (t/\theta)^n dt) = (1+1/n)(\theta - \theta^{-n}\theta^{n+1}/(n+1)) \\
&= (1+1/n)(\theta - \theta/(n+1)) = \theta\frac{n+1}{n}\frac{n}{n+1} = \theta.
\end{aligned}$$

From Definition 1.37, $\operatorname{var}((1+1/n)X_{(n)})$ is equal to

$$\begin{aligned}
&\frac{(n+1)^2}{n^2}\mathbf{E}X_{(n)}^2 - \theta^2 = \frac{(n+1)^2}{n^2}\int_0^\theta 2t\mathbf{P}(X_{(n)} > t)dt - \theta^2 \\
&= \theta^2\Big(\frac{(n+1)^2}{n^2} - 1\Big) - \frac{(n+1)^2}{n^2}\int_0^\theta 2t\mathbf{P}(X_{(n)} < t)dt \\
&= \theta^2\Big(\frac{(n+1)^2}{n^2} - 1\Big) - \frac{(n+1)^2}{n^2}\theta^{-n}\int_0^\theta 2tt^n dt = \theta^2\Big(\frac{(n+1)^2}{n^2} - 1\Big) - \frac{(n+1)^2}{n^2}\theta^2\frac{2}{n+2} \\
&= \frac{\theta^2}{n^2(n+2)}\Big((n+1)^2(n+2) - n^2(n+2) - 2(n+1)^2\Big) \\
&= \frac{\theta^2}{n^2(n+2)}\Big([(n+1)^2 - n^2](n+2) - 2(n+1)^2\Big) \\
&= \frac{\theta^2}{n^2(n+2)}\Big([2n+1](n+2) - 2(n+1)^2\Big) = \frac{\theta^2}{n^2(n+2)}(5n - 4n + 2 - 2) = \frac{\theta^2}{n(n+2)}.
\end{aligned}$$

In fact, $(1+1/n)X_{(n)}$ is the uniform minimum variance unbiased estimator for θ (and we call this estimator UMVU), though we will not prove it.

Example 4.8. Suppose we have a binomial random variable with unknown parameters n, p. We want to find method of moments estimators for n and for 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

$$\mu_1 = np, \qquad \mu_2 = np(1-p) + n^2p^2,$$

to get an estimator for n:

$$N := \frac{M_1^2}{M_1 - (M_2 - M_1^2)}, \qquad \text{since} \qquad n = \frac{\mu_1^2}{\mu_1 - (\mu_2 - \mu_1^2)},$$

and an estimator for p:

$$P := \frac{M_1}{N}, \qquad \text{since} \qquad p = \frac{\mu_1}{n}.$$

(To solve the system, note that the second equation says $\mu_2 = (1-p)\mu_1 + \mu_1^2 = (1-\mu_1/n)\mu_1 + \mu_1^2$, and then solve for n.)

4.2. **Sufficient Statistics.** Suppose we have some data and a family of distributions $\{f_\theta : \theta \in \Theta\}$. We would like to find the parameter θ among the distributions that fits the data well. One way to achieve this goal is to look for a sufficient statistic. Once we find the sufficient statistic, we can then apply the Rao-Blackwell Theorem, Theorem 4.17 below, to get a good estimate of the parameter θ.

Definition 4.9 (**Sufficient Statistic**)**.** Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of densities. Let $t : \mathbb{R}^n \to \mathbb{R}^k$, 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 4.10. 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 4.11. 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

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

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

Theorem 4.12 (**Factorization Theorem**). *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 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 $Y := t(X_1,\ldots,X_n)$ 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*

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

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 $B_1, B_2,\ldots$ such that the total volume of $B_1, B_2,\ldots$ 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.

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. Define $r_\theta(z) := \mathbf{P}_\theta(t(X)=z)$ for any $z\in\mathbb{R}^k$. 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

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

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

Remark 4.13. If $t(x) := x$ for all $x \in \mathbb{R}^n$, then the statistic $t(X_1, \ldots, X_n) = (X_1, \ldots, X_n)$ is automatically sufficient for θ, choosing $g_\theta := f_\theta$ and $h := 1$. So, at least one sufficient statistic always exists.

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. So, using the entire data set $(X_1, \ldots, X_n)$ as a sufficient statistic is in some sense undesirable. It would be nice to have a sufficient statistic with a more succinct representation. This goal can be realized by defining a minimal sufficient statistic, but this concept is beyond the scope of this course.

4.3. **Evaluating Estimators.**

Exercise 4.14 (**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).)

Even if an estimator is unbiased, its distribution of values might be quite far from $g(\theta)$. 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 $g(\theta)$:

$$\mathbf{E}_\theta(Y - g(\theta))^2.$$

If the estimator is unbiased, this quantity is equal to the variance of Y.

Definition 4.15 (**UMVU**)**.** Let $X_1, \ldots, X_n$ 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 $Y := t(X_1, \ldots, X_n)$ 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

$$\mathrm{Var}_\theta(Y) \leq \mathrm{Var}_\theta(Z), \qquad \forall\, \theta \in \Theta.$$

Remark 4.16. 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.

The Rao-Blackwell Theorem says that any sufficient statistic can be used to improve any estimator for $g(\theta)$.

Theorem 4.17 (**Rao-Blackwell**)**.** *Let Z be a sufficient statistic for $\{f_\theta \colon \theta \in \Theta\}$ and let Y be an unbiased estimator for θ. Define $W := \mathbf{E}_\theta(Y|Z)$. (Since Z is sufficient for θ, W does not depend on θ by Exercise 4.20, i.e. W is a well-defined function of the random sample but not an explicit function of θ.) Let $\theta \in \Theta$ with $\mathrm{Var}_\theta(Y) < \infty$. Then*

$$\mathrm{Var}_\theta(W) \leq \mathrm{Var}_\theta(Y).$$

This inequality is strict unless $W = Y$.

Proof. By the (conditional) Jensen's inequality, Exercise 4.19

$$(W - \theta)^2 = (\mathbf{E}_\theta(Y|Z) - \theta)^2 \leq \mathbf{E}_\theta[(Y-\theta)^2|Z].$$

Taking expected values of both sides and applying Exercise 4.14(ii), we get

$$\mathrm{Var}_\theta(W) \leq \mathbf{E}_\theta\mathbf{E}_\theta[(Y-\theta)^2|Z] = \mathbf{E}_\theta(Y-\theta)^2 = \mathrm{Var}_\theta(Y).$$

And 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 $W = Y$. □

Recall that the function $t \mapsto t^2$ is a convex function of $t \in \mathbb{R}$.

Definition 4.18. 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.

Exercise 4.19 (**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 7.10) 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 4.20. 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 4.21. By Exercise 4.14, 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 4.17.

Remark 4.22. 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 4.23. 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 4.17. 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 4.14 (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 $1 \leq k < \ell \leq n$, 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 4.14,

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

Therefore, by Exercise 4.14(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 4.17 did in fact substantially improve our estimator $Y = X_1$, since W has variance of order n^{-1}, while Y has constant variance.

4.4. **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 4.24 (**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

$$I(\theta) = I_X(\theta) := \mathbf{E}_\theta(\frac{d}{d\theta} \log f_\theta(X))^2, \qquad \forall\, \theta \in \Theta,$$

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 8.8 applies.) We also have another equivalent definition:

$$\begin{aligned} \mathbf{E}_\theta \frac{d^2}{d\theta^2} \log f_\theta(X) &= \int_{\mathbb{R}^n} \frac{d}{d\theta} \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) dx = \int_{\mathbb{R}^n} \frac{f_\theta(x) \frac{d^2}{d\theta^2} f_\theta(x) - \Big(\frac{d}{d\theta} f_\theta(x)\Big)^2}{[f_\theta(x)]^2} f_\theta(x) dx \\ &= \int_{\mathbb{R}^n} \frac{d^2}{d\theta^2} f_\theta(x) - \Big(\frac{d}{d\theta} \log f_\theta(x)\Big)^2 f_\theta(x) dx = 0 - I_X(\theta) = -I_X(\theta). \end{aligned}$$

The Fisher information expresses the amount of "information" a random variable has.

Example 4.25. 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 4.26. *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 4.27. 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 $\mathbf{P}(X = x, Y = y)/\mathbf{P}(Y = y)$.)

Our primary interest in information is the following inequality. Theorem 4.28 gives a lower bound on the variance of unbiased estimators of θ.

Theorem 4.28 (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.

(Differentiation under the integral sign in the proof can be justified whenever Proposition 8.8 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.)

Remark 4.29. In the case that $X_1, \ldots, X_n$ are i.i.d. real-valued random variables and $X = (X_1, \ldots, X_n)$, Proposition 4.26 says that $I_X(\theta) = \sum_{i=1}^n I_{X_i}(\theta) = nI_{X_1}(\theta)$. And if Y is unbiased for θ, Theorem 4.28 says

$$\mathrm{Var}_\theta(Y) \geq \frac{1}{nI_{X_1}(\theta)}, \qquad \forall\, \theta \in \Theta.$$

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,

$$|g'(\theta)| = \left| \frac{d}{d\theta} \int_{\mathbb{R}^n} f_\theta(x) t(x) dx \right| = \left| \int_{\mathbb{R}^n} \frac{d}{d\theta} \log f_\theta(x) t(x) f_\theta(x) dx \right| = \left| \mathbf{E}_\theta \frac{d}{d\theta} \log f_\theta(X) t(X) \right|$$
$$= \left| \mathrm{Cov}_\theta\left(\frac{d}{d\theta} \log f_\theta(X), t(X)\right) \right| \leq \sqrt{\mathrm{Var}_\theta\left(\frac{d}{d\theta} \log f_\theta(X)\right) \mathrm{Var}_\theta(t(X))} = \sqrt{I_X(\theta) \mathrm{Var}_\theta(t(X))}.$$

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 4.28 allows us to find a UMVU for θ. To find such an estimator, we look for affine functions of $\frac{d}{d\theta} \log f_\theta(X)$.

Example 4.30. 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

$$\frac{d}{d\theta} \log \prod_{i=1}^n f_\theta(x_i) = \sum_{i=1}^n \left(\frac{1}{\theta} + \log x_i\right) = n\left(\frac{1}{\theta} + \frac{1}{n} \log \prod_{i=1}^n x_i\right), \qquad \forall\, 0 < x_i < 1,\ 1 \leq i \leq n.$$

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 4.24, 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 4.28, Y must be UMVU for $1/\theta = \mathbf{E}_\theta Y$.

Theorem 4.28 suggests the following quantity represents the efficiency of an estimator.

Definition 4.31 (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

$$\frac{I_X(\theta) \mathrm{Var}_\theta(Z)}{I_X(\theta) \mathrm{Var}_\theta(Y)} = \frac{\mathrm{Var}_\theta(Z)}{\mathrm{Var}_\theta(Y)}, \qquad \forall\, \theta \in \Theta.$$

4.5. **Maximum Likelihood Estimator.** Let $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ as

$$\prod_{i=1}^{n} f_\theta(x_i), \qquad \forall\, 1 \le i \le n.$$

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 4.32 (**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 $t(x_1, \dots, x_n)$ 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 4.33. 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 4.34. Let $f_\theta(x_1) := 1_{[\theta, \theta+1]}(x_1)$ for all $x_1, \theta \in \mathbb{R}$. Then, for all $x_1, \dots, x_n, \theta \in \mathbb{R}$, 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 $\phi \colon \mathbb{R}^n \to \mathbb{R}$ is convex if for any $x, y \in \mathbb{R}^n$ with $x \neq y$ and for any $t \in (0, 1)$,

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

And $\phi \colon \mathbb{R}^n \to \mathbb{R}$ 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 4.35 (**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 4.34 shows that log-concavity still does not guarantee uniqueness of the maximum of the likelihood function. However, strict log-concavity does guarantee uniqueness.

Proposition 4.36. *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 $1 \le i \le n$. Fix $x_1,\dots,x_n \in \mathbb{R}$. 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 4.39. From Exercise 4.37, this function has at most one global maximum. □

Exercise 4.37. 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 4.38. 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 4.39. 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 4.40. 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

$$f(x_1, z) < f(x_1, x_2), \qquad \forall\, x_2 \ne z.$$

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 4.41. 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 1.96 that, if $z \neq \frac{1}{n}\sum_{i=1}^{n} x_i$, then

$$\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 < \frac{1}{n}\sum_{i=1}^{n}(x_i - z)^2.$$

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 M_1 is UMVU for the mean (by Example 4.11 M_1 is sufficient for the mean, by the Rao-Blackwell Theorem 4.17 $\mathbf{E}_\theta(M_1|M_1)$ is UMVU for the mean, and $\mathbf{E}_\theta(M_1|M_1) = M_1$ by Exercise 4.14(iv)). 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 3.7 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. If we fix $\mu \in \mathbb{R}$ and look at the information of the n-dimensional Gaussian X, we get by modifying Example 4.25 and using Proposition 4.26

$$\begin{aligned} I_X(\sigma) = nI_{X_1}(\sigma) &= n\mathrm{Var}_\sigma\Big(\frac{d}{d\sigma}\frac{-(X_1-\mu)^2}{2\sigma^2}\Big) = n\sigma^{-6}\mathrm{Var}_\sigma[(X_1-\mu)^2] \\ &= n\sigma^{-6}\mathbf{E}_\sigma((X_1-\mu)^4 - \sigma^4) = 2n\sigma^{-2}. \end{aligned}$$

By the Cramér-Rao Inequality, Theorem 4.28, with $g(\sigma) = \mathbf{E}_\sigma(Y) = \sigma^2(n-1)/n$ (using Proposition 3.7), 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 3.7,

$$\mathrm{Var}_\sigma(Y) = \mathrm{Var}_\sigma\Big[\frac{\sigma^2}{n}\frac{1}{\sigma^2}\sum_{j=1}^n\Big(X_j - \frac{1}{n}\sum_{i=1}^n X_i\Big)^2\Big] = \frac{\sigma^4}{n^2}2(n-1) = \frac{2\sigma^4(n-1)}{n^2}.$$

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 4.42. 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 4.43. Consider a random sample from the exponential density $1_{x>0}\theta e^{-\theta x}$ with $\theta > 0$ unknown. Then

$$\log\prod_{i=1}^n 1_{x_i>0}\theta e^{-\theta x_i} = 1_{x_1,\ldots,x_n>0}\log\theta - \theta\sum_{i=1}^n x_i.$$

So,

$$\frac{d}{d\theta}\log\prod_{i=1}^n 1_{x_i>0}\theta e^{-\theta x_i} = 1_{x_1,\ldots,x_n>0}\frac{n}{\theta} - \sum_{i=1}^n x_i.$$

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 3.13, 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 with variance $(g'(1/\theta))^2\theta^{-2} = \theta^2$ as $n \to \infty$. That is, (using also Theorem 3.15)

$$\operatorname{Var}(Y) = \operatorname{Var}\Big[n^{-1/2}\sqrt{n}\Big(\frac{1}{\overline{X}_n} - \theta\Big)\Big] = \frac{1}{n}\theta^2(1 + o(1)).$$

On the other hand, the information inequality, Theorem 4.28, 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 4.44. 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 4.45 generalizes this observation.

Proposition 4.45 (**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.

Note that MoM estimators also satisfy Functional Equivariance.

Lemma 4.46 (**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 4.47. 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 4.48 (Consistency of MLE). *Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}^n$ be i.i.d. random variables with common probability density $f_\theta\colon \mathbb{R}^n \to [0,\infty)$. Fix $\theta \in \Theta \subseteq \mathbb{R}^m$. Suppose Θ is compact and $f_\theta(x_1)$ is a continuous function of θ for a.e. $x_1 \in \mathbb{R}^n$. (Then the maximum of $\ell(\theta)$ exists, since it is a continuous function on a compact set.) 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 $n \to \infty$, the 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 4.46 (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 bijection, it follows from Proposition 4.45 that the MLE for $g(\theta)$ is also consistent.

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 4.49 (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 $\Theta \subseteq \mathbb{R}$. Assume the following*

(i) *The set $A := \{x \in \mathbb{R}\colon f_\theta(x) > 0\}$ 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*

$$\mathbf{E}_\theta \sup_{\theta' \in \Theta} \left| 1_{\theta' \in [\theta-\varepsilon, \theta+\varepsilon]} \frac{d^2}{d[\theta']^2} \log f_{\theta'}(X_1) \right| < \infty.$$

(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 with variance $\frac{1}{I_{X_1}(\theta)}$, with respect to $\mathbf{P}_\theta$.

Remark 4.50. Combining this Theorem with Proposition 4.45, under the above assumptions (and also if the variance of the MLE converges, i.e. we can apply something like Theorem 3.15), the MLE for θ achieves the asymptotically optimal variance in the Cramér-Rao Inequality, Theorem 4.28. The same holds for an 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

$$\sqrt{n}(Y_n - \theta) = \frac{\sqrt{n}\ell_n'(\theta)}{-\ell_n''(Z_n)}, \qquad \text{if } A_n \text{ occurs.} \qquad (*)$$

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 4.51. □

Exercise 4.51. 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 $A_1, A_2, \ldots \subseteq \Omega$ satisfy $\lim_{n\to\infty} \mathbf{P}(A_n) = 1$. Then, as $n \to \infty$

$$W_n 1_{A_n} + U_n 1_{A_n^c}$$

converges in distribution to W.

4.6. **Additional Comments.** The Rao-Blackwell Theorem also applies for convex loss functions other than the squared difference:

Exercise 4.52 (**Rao-Blackwell, Generalized**)**.** Suppose we are given a **loss function**

$$\ell\colon \Theta \times \mathbb{R}^k \to \mathbb{R},$$

and we are 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 $y, \theta \in \mathbb{R}$.

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 4.20, 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 $y \in \mathbb{R}$. Show that

$$r(\theta, W) \leq r(\theta, Y).$$

And if $\ell(\theta, y)$ is strictly convex in y, then this inequality is strict unless $W = Y$.

The Cramér-Rao and Limiting Distribution for the MLE have analogous statements when Θ is a vector space.

Theorem 4.53 (**Multiparameter Cramér-Rao/ Information Inequality**)**.** *Suppose $X\colon \Omega \to \mathbb{R}^n$ is 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 8.8 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 vector of variances of the components of Y. Then

$$\mathrm{Var}_\theta(Y) \geq (Dg(\theta))^T [I_X(\theta)]^{-1} Dg(\theta), \qquad \forall\, \theta \in \Theta.$$

In particular, if Y is unbiased for θ,

$$\mathrm{Var}_\theta(Y) \geq [I_X(\theta)]^{-1}, \qquad \forall\, \theta \in \Theta.$$

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.

Theorem 4.54 (**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 $x \in A$, $\forall\ 1 \le i, j \le m$, $\frac{\partial^2 f_\theta(x)}{\partial\theta_i \partial\theta_j}$ 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, $\forall\ 1 \le i, j \le m$, $\exists\ \varepsilon > 0$ such that*

$$\mathbf{E}_\theta \sup_{\theta' \in \Theta} \left| 1_{\theta' \in [\theta-\varepsilon, \theta+\varepsilon]} \frac{\partial^2}{\partial\theta_i' \partial\theta_j'} \log f_{\theta'}(X_1) \right| < \infty.$$

(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 $\mathbf{P}_\theta$.

5. Hypothesis Testing

In Section 4, we gave methods for estimating parameters from a probability distribution with unknown parameters. In this section, we consider the corresponding "decision problem" for parameter estimation. For example, we consider whether or not an unknown parameter lies in a certain range of values, and we try to estimate the probability of this event. A hypothesis is then a guess for the value or range of values of an unknown parameter.

Definition 5.1 (**Null Hypothesis, Alternative Hypothesis**). Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of distributions. Let $\Theta_0 \subseteq \Theta$. A **null hypothesis** H_0 is an event of the form

$$\{\theta \in \Theta_0\}.$$

Define $\Theta_1 := \Theta_0^c$, so that $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$. The **alternative hypothesis** H_1 is the event

$$\{\theta \in \Theta_1\}.$$

Example 5.2. In Exercise 2.19, we supposed that we had a roulette wheel such that, with probability p, red results from one spin of the roulette wheel. So we can take $\Theta = [0,1]$, H_0 to be the event $\{\theta = 18/38\}$, and H_1 is the event $\{\theta \in [0,1] \colon \theta \neq 18/38\}$.

Example 5.3. Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of distributions. Suppose Θ, Θ_0 are such that $\{f_\theta \colon \theta \in \Theta_0\}$ is the set of all Gaussian densities with unknown mean and variance, and $\{f_\theta \colon \theta \in \Theta_1\}$ is some other set of non-Gaussian probability density functions. Then the null-hypothesis H_0 is the assertion that f_θ is a Gaussian density (with arbitrary mean and variance), and the alternative hypothesis H_1 is the assertion that f_θ is in the remaining set of probability densities.

5.1. **Neyman-Pearson Testing.** Let $X\colon \Omega \to \mathbb{R}^n$ be a random variable with distribution f_θ, where $\{f_\theta\colon \theta \in \Theta\}$ is a family of multivariable probability densities or probability mass functions.

Definition 5.4 (**Critical Region/ Rejection Region**)**.** Let H_0 be a null hypothesis. A **hypothesis test** of H_0 versus H_1 is specified by a subset $C \subseteq \mathbb{R}^n$. The set C is called the **critical region** or the **rejection region**. The test proceeds as follows:

- If $X \notin C$, then we accept the null hypothesis H_0 to be true.
- If $X \in C$, then we reject the null hypothesis H_0, and instead assert that H_1 is true.

The region $C^c \subseteq \mathbb{R}^n$ is called the **acceptance region**. The performance of the test is quantified by its **power function** $\beta\colon \Theta \to [0,1]$ defined by

$$\beta(\theta) := \mathbf{P}_\theta(X \in C) = 1 - \mathbf{P}_\theta(X \notin C), \qquad \forall\, \theta \in \Theta.$$

WARNING: This notation of power is different than that used in the Rice book.

In an ideal world, we could find a test that performs perfectly, i.e. we would prefer that $\beta(\theta) = 0$ for all $\theta \in \Theta_0$ and $\beta(\theta) = 1$ for all $\theta \in \Theta_1$. In practice, such a β often cannot be found. For example, H_0 may be accepted to be true while actually being false.

Definition 5.5 (**Type II Error**)**.** A **Type II Error** for a hypothesis test occurs when $X \notin C$ with positive probability, but H_0 is actually false. That is, $\beta(\theta) < 1$ for some $\theta \in \Theta_1$. That is, H_0 is accepted to be true by the test, while actually being false.

The quantity $1 - \beta(\theta)$ is the probability of occurrence of a Type II Error for $\theta \in \Theta_1$.

WARNING: This notation for the probability of a Type II Error is different than that used in the Rice book.

A type II error is sometimes called a "false negative."

It is also undesirable that H_1 may be accepted to be true while actually being false.

Definition 5.6 (**Type I Error**)**.** A **Type I Error** for a hypothesis test occurs when $X \in C$ with positive probability, but H_1 is actually false (i.e. H_0 is true). That is, $\beta(\theta) > 0$ for some $\theta \in \Theta_0$. That is, H_1 is accepted to be true by the test, while actually being false.

The value of $\beta(\theta)$ is the probability of occurrence of a Type I Error for $\theta \in \Theta_0$.

The **significance level** α is defined as

$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta).$$

A type I error is sometimes called a "false positive." So, α is the "worst" probability of a false positive occurring.

Example 5.7. Let us return to Exercise 2.19. The roulette wheel has 38 spaces and 18 red spaces. Suppose we spin the roulette wheel 5 times resulting in X red outcomes. We model the set of outcomes as a sum of independent $\{0,1\}$ valued random variables, so that the total number of red outcomes X is a binomial random variable with parameters n, θ with $n = 5$ and $\theta \in [0,1]$ unknown. Suppose the null hypothesis H_0 is $\{0 \le \theta \le 1/2\}$, and the alternative hypothesis H_1 is $\{1/2 < \theta \le 1\}$. If θ is small, then the observed value of X should be small as well, so a "good" hypothesis test should use a rejection region consisting of large values of X.

Recalling that $0 \leq X \leq 5$, let's first consider a test for this hypothesis that rejects H_0 if and only if $X = 5$. That is, $C := \{5\}$, and

$$\beta(\theta) = \mathbf{P}_\theta(X \in C) = \mathbf{P}_\theta(X = 5) = \theta^5.$$

For this test, the probability of a type I error is fairly low since it is at most

$$\alpha = \sup_{\theta \in [0,1/2]} \beta(\theta) = \beta(1/2) = (1/2)^5 \approx .03.$$

However, the probability of a type II error is quite far from 0, since e.g. $1 - \beta(.6) \approx .92$, and $1 - \beta(.87) \approx .5$.

Let us therefore consider a different test that improves on the type II error. Suppose we now reject H_0 if and only if $X \in \{3, 4, 5\}$. That is, $C := \{3, 4, 5\}$, and

$$\begin{aligned}\beta(\theta) &= \mathbf{P}_\theta(X \in C) = \mathbf{P}_\theta(X = 3 \text{ or } X = 4 \text{ or } X = 5) \\ &= \binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta)^4 + \binom{5}{5}\theta^5.\end{aligned}$$

For this test, the probability of a type I error is not quite as good:

$$\alpha = \sup_{\theta \in [0,1/2]} \beta(\theta) = \beta(1/2) = 1/2.$$

However, the probability of a type II error is better than before, since e.g. $1 - \beta(.6) \approx .32$, and $1 - \beta(.87) \approx .017$.

From the above example, we see that different tests can have different Type I and Type II Errors, and it might be a priori unclear which test is the "best." In practice, one fixes some bound on the significance level α such as $\alpha = .05$. For example, in driverless cars, the autonomous system constantly tests the hypothesis H_0 that "there is an obstruction ahead of the car such that the brakes need to be applied." We would like to have a small upper bound on α, since a Type I Error corresponds to an obstruction being present, but the autonomous system does not believe this to be the case (so the car does not apply the brakes). In this example, a Type II Error corresponds to the car applying the brakes unnecessarily, which is also undesirable but perhaps less so than a Type I error.

Definition 5.8 (**Uniformly Most Powerful Test (UMP)**). Let $\Theta_0 \subseteq \Theta$ and denote $\Theta_1 := \Theta_0^c$. Let H_0 be the hypothesis $\{\theta \in \Theta_0\}$ and let H_1 be the hypothesis $\{\theta \in \Theta_1\}$. Let $\mathcal{T}$ be a family of hypothesis tests. A hypothesis test in $\mathcal{T}$ with power function $\beta(\theta)$ is called **Uniformly Most Powerful (UMP) class $\mathcal{T}$ test** if

$$\beta(\theta) \geq \beta'(\theta), \qquad \forall\, \theta \in \Theta_1,$$

for every $\beta'(\theta)$ that is a power function of any hypothesis test in $\mathcal{T}$.

In the case that Θ consists of exactly two points, it is possible to explicitly find a UMP among all hypothesis tests with significance level at most α, where $\alpha \in [0, 1]$. This UMP is given by a likelihood ratio test.

Lemma 5.9 (**Neyman-Pearson**). *Suppose $\Theta = \{\theta_0, \theta_1\}$, $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$. Let H_0 be the hypothesis $\{\theta = \theta_0\}$ and let H_1 be the hypothesis $\{\theta = \theta_1\}$. Let $\{f_{\theta_0}, f_{\theta_1}\}$ be*

two multivariable probability densities or probability mass functions. Fix $k \geq 0$. Define a ***likelihood ratio test*** *with rejection region C in the following way:*

$$C = \{x \in \mathbb{R}^n \colon f_{\theta_1}(x) \geq k f_{\theta_0}(x)\}. \qquad (*)$$

Define

$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0) = \mathbf{P}_{\theta_0}(X \in C). \qquad (**)$$

Let $\mathcal{T}$ be the class of all tests with significance level at most α. Then

- *(Sufficiency) Any hypothesis test satisfying $(*)$ is a UMP class $\mathcal{T}$ test.*
- *(Necessity) If there exists a hypothesis test satisfying $(*)$ and $(**)$ with $k > 0$, then any UMP class $\mathcal{T}$ test has significance level equal to α, and any UMP class $\mathcal{T}$ test satisfies $(*)$, except possibly on a set $D \subseteq \mathbb{R}^n$ satisfying $\mathbf{P}_{\theta_0}(X \in D) = \mathbf{P}_{\theta_1}(X \in D) = 0$.*

Remark 5.10. Intuitively, the likelihood ratio test compares how "likely" $x \in \mathbb{R}^n$ is to satisfy the null hypothesis (quantified by $f_{\theta_0}(x)$) to how "likely" $x \in \mathbb{R}^n$ is to satisfy the alternative hypothesis (quantified by $f_{\theta_1}(x)$). If the null hypothesis is not very "likely" to occur, i.e. $f_{\theta_0}(x)$ is a bit smaller than $f_{\theta_1}(x)$, then the test rejects the null hypothesis.

Remark 5.11. When we use Lemma 5.9 in practice, we will typically fix $\alpha \in [0, 1]$, and then define $k \geq 0$ such that $\mathbf{P}_{\theta_0}(X \in C)$ is at most α (even though Lemma 5.9 instead starts with C and then defines α via C).

Proof. In the proof below we assume that $\{f_{\theta_0}, f_{\theta_1}\}$ are multivariable probability densities. The probability mass function case follows by replacing the integral below by a sum.

As we already noted in $(**)$, Θ_0 consists of a single point, so the supremum appearing in $(**)$ is just $\beta(\theta_0)$, and we will repeatedly use this fact below without further mention.

Let $\beta(\theta)$ be the power function of the test corresponding to C. Let C' be the rejection region of any class $\mathcal{T}$ test, and let $\beta'(\theta)$ be the power function of this test. By definition of C, we have

$$[1_C(x) - 1_{C'}(x)][f_{\theta_1}(x) - k f_{\theta_0}(x)] \geq 0, \qquad \forall\, x \in \mathbb{R}^n.$$

Therefore,

$$\begin{aligned} 0 &\leq \int_{\mathbb{R}^n} [1_C(x) - 1_{C'}(x)][f_{\theta_1}(x) - k f_{\theta_0}(x)] dx \\ &= \mathbf{P}_{\theta_1}(X \in C) - \mathbf{P}_{\theta_1}(X \in C') - k[\mathbf{P}_{\theta_0}(X \in C) - \mathbf{P}_{\theta_0}(X \in C')] \\ &= \beta(\theta_1) - \beta'(\theta_1) - k[\beta(\theta_0) - \beta'(\theta_0)]. \qquad (***) \end{aligned}$$

Since the test corresponding to C has significance level α and the test corresponding to C' has significance level at most α, we have $\beta(\theta_0) - \beta'(\theta_0) \geq 0$. So, $k \geq 0$ and $(***)$ imply that $\beta(\theta_1) - \beta'(\theta_1) \geq 0$. That is, the C test is UMP class $\mathcal{T}$.

We now prove necessity. Let C' be the rejection region corresponding to a UMP class $\mathcal{T}$ test. We just showed that the C test is UMP class $\mathcal{T}$. Therefore $\beta(\theta_1) = \beta'(\theta_1)$. Using this fact, $(***)$ and $k > 0$, we then get

$$\alpha - \beta'(\theta_0) \overset{(**)}{=} \beta(\theta_0) - \beta'(\theta_0) \overset{(***)}{\leq} 0. \qquad (\ddagger)$$

Since C' is a UMP class $\mathcal{T}$ test, the C' test has significance level at most α, i.e. $\beta'(\theta_0) \leq \alpha$, so that $\beta'(\theta_0) = \alpha$ by $(\ddagger)$. So, $(***)$ is equal to zero, and the nonnegative integrand appearing in $(***)$ must be equal to zero. The final assertion follows. $\square$

Example 5.12. Suppose X is a binomial distributed random variable with parameters 2 and $\theta \in \{1/2, 3/4\}$. We want to test the hypothesis H_0 that $\theta = 1/2$ versus the hypothesis H_1 that $\theta = 3/4$. Lemma 5.9 says that the UMP test for the class of tests with an upper bound on the significance level must be a likelihood ratio test. There are only three values that X can take, so we examine the likelihood ratios explicitly:

$$\frac{f_{3/4}(0)}{f_{1/2}(0)} = \frac{(1-3/4)^2}{(1-1/2)^2} = \frac{1}{4}, \qquad \frac{f_{3/4}(1)}{f_{1/2}(1)} = \frac{2(1-3/4)(3/4)}{2(1-1/2)(1/2)} = \frac{3}{4}, \qquad \frac{f_{3/4}(2)}{f_{1/2}(2)} = \frac{(3/4)^2}{(1/2)^2} = \frac{9}{4}.$$

We then get different likelihood ratio tests according to the choice of $k > 0$.

- If $3/4 < k \leq 9/4$, then H_0 is rejected if and only if $X = 2$, and this test is the unique UMP for tests with significance level at most $\mathbf{P}_{1/2}(X = 2) = 1/4$.
- If $1/4 < k \leq 3/4$, then H_0 is rejected if and only if $X = 1$ or 2, and this test is the unique UMP for tests with significance level at most $\mathbf{P}_{1/2}(X \in \{1, 2\}) = 3/4$.
- If $0 < k \leq 1/4$, then H_0 is always rejected, and this test is the unique UMP for tests with significance level at most $\mathbf{P}_{1/2}(X \in \{1, 2, 3\}) = 1$.
- If $k > 9/4$, then H_0 is never rejected, and this test is the unique UMP for tests with significance level at most $\mathbf{P}_{1/2}(X \in \emptyset) = 0$.

Note that $\mathbf{P}_{1/2}(X \in \{0, 1, 2\}) = \mathbf{P}_{3/4}(X \in \{0, 1, 2\}) = 1$, so we do not need to consider the last part of Lemma 5.9.

Exercise 5.13. Suppose X is a Gaussian distributed random variable with known variance $\sigma^2 > 0$ but unknown mean. Fix $\mu_0, \mu_1 \in \mathbb{R}$. Assume that $\mu_0 - \mu_1 > 0$. We want to test the hypothesis H_0 that $\mu = \mu_0$ versus the hypothesis H_1 that $\mu = \mu_1$. Fix $\alpha \in (0, 1)$. Explicitly describe the UMP test for the class of tests whose significance level is at most α.

Your description of the test should use the function $\Phi(t) := \int_{-\infty}^{t} e^{-x^2/2} dx/\sqrt{2\pi}$, $\Phi\colon \mathbb{R} \to (0, 1)$, and/or the function $\Phi^{-1}\colon (0, 1) \to \mathbb{R}$. (Recall that $\Phi(\Phi^{-1}(s)) = s$ for all $s \in (0, 1)$ and $\Phi^{-1}(\Phi(t)) = t$ for all $t \in \mathbb{R}$.)

5.2. Hypothesis Tests and Confidence Intervals.

Definition 5.14 (Confidence Interval, Confidence Region). Let $X\colon \Omega \to \mathbb{R}^n$ be a random variable with distribution f_θ, where $\{f_\theta\colon \theta \in \Theta\}$ is a family of multivariable probability densities or probability mass functions. Let $g\colon \Theta \to \mathbb{R}$. Let $u, v\colon \mathbb{R}^n \to \mathbb{R}$ such that $u(x) \leq v(x)$ for all $x \in \mathbb{R}^n$. Let $\alpha \in (0, 1)$. A **100(1-α)% confidence interval** for a parameter $g(\theta)$ is a random interval of the form $[u(X), v(X)]$ satisfying

$$\mathbf{P}_\theta(g(\theta) \in [u(X), v(X)]) \geq 1 - \alpha, \qquad \forall\, \theta \in \Theta.$$

More generally, if $c\colon \mathbb{R}^n \to 2^\Theta$, then a **100(1-$\alpha$)% confidence region** for a parameter $g(\theta)$ is a random set $c(X)$ satisfying

$$\mathbf{P}_\theta(g(\theta) \in c(X)) \geq 1 - \alpha, \qquad \forall\, \theta \in \Theta.$$

Example 5.15. Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $[0, 1]$ with unknown mean $\mu \in [0, 1]$ and known variance $\sigma^2 \in (0, 1)$. Let $X := \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean. Then $\mathbf{E}X = \mu$ and $\mathrm{Var}(X) = \frac{\sigma^2}{n}$. From the Central Limit Theorem with error bound (i.e. the Berry-Esseén Theorem 2.30),

$$\sup_{t \in \mathbb{R}} \left| \mathbf{P}\Big(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < t \Big) - \mathbf{P}(Z < t) \right| \leq \frac{1}{\sigma^3 \sqrt{n}}.$$

Choosing e.g. $t = 2$ and $t = -2$ and subtracting the results,

$$\left|\mathbf{P}\Big(-2 < \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < 2\Big) - \mathbf{P}(-2 < Z < 2)\right| \leq \frac{2}{\sigma^3\sqrt{n}}$$

That is, we get a confidence interval for the parameter μ for any $n \geq 1$:

$$\mathbf{P}\Big(\frac{X_1 + \cdots + X_n}{n} - 2\frac{\sigma}{\sqrt{n}} < \mu < \frac{X_1 + \cdots + X_n}{n} + 2\frac{\sigma}{\sqrt{n}}\Big) \geq \mathbf{P}(-2 < Z < 2) - \frac{2}{\sigma^3\sqrt{n}} \geq .95 - \frac{2}{\sigma^3\sqrt{n}}.$$

There is a straightforward duality between hypothesis tests and confidence regions.

Proposition 5.16 (**Confidence Region/ Hypothesis Test Duality**). *Let $X\colon \Omega \to \mathbb{R}^n$ be a random variable.*

- *Fix $\alpha \in (0,1)$. Assume that for every $\theta_0 \in \Theta$, there is a hypothesis test with significance level α of the hypothesis H_0 that is $\{\theta = \theta_0\}$. Let $C(\theta_0) \subseteq \mathbb{R}^n$ denote the rejection region of this test. Then the set*

  $$c(X) := \{\theta \in \Theta\colon X \notin C(\theta)\}$$

  is a $100(1-\alpha)\%$ confidence region for θ.
- *Let $c\colon \mathbb{R}^n \to 2^\Theta$. Assume that $c(X)$ is a $100(1-\alpha)\%$ confidence region for θ. Define a hypothesis test of $\theta = \theta_0$ whose rejection region is*

  $$C(\theta) := \{x \in \mathbb{R}^n\colon \theta \notin c(x)\}.$$

  Then this test has significance level at most α.

Proof. For the first statement, note that $\alpha = \sup_{\theta\in\Theta_0}\beta(\theta) = \beta(\theta_0) = \mathbf{P}_{\theta_0}(X \in C(\theta_0))$. By the definition of $c(X)$ and $C(\theta)$, for any $\theta \in \Theta$,

$$\mathbf{P}_\theta(\theta \in c(X)) = \mathbf{P}_\theta(X \notin C(\theta)) = 1 - \alpha.$$

The first statement follows. For the second statement, the definition of $c(X)$ and $C(\theta)$ gives

$$1 - \alpha \leq \mathbf{P}_\theta(\theta \in c(X)) = \mathbf{P}_\theta(X \notin C(\theta)) = 1 - \mathbf{P}_\theta(X \in C(\theta)), \qquad \forall\, \theta \in \Theta.$$

The second statement then follows, since $\sup_{\theta\in\Theta_0}\beta(\theta) = \beta(\theta_0) = \mathbf{P}_{\theta_0}(X \in C(\theta_0)) \leq \alpha$. □

5.3. **p-Value.** A p-value is a measure of the belief of rejecting the null hypothesis. A small p-value corresponds to a high probability that the null hypothesis is false.

Definition 5.17 (**p-Value**). Let $X_1, \ldots, X_n$ be a real-valued random sample of size n from a family of distributions $\{f_\theta\colon \theta \in \Theta\}$. Denote $X := (X_1, \ldots, X_n)$. Let $t\colon \mathbb{R}^n \to \mathbb{R}$. Let $Y := t(X)$. For any $c \in \mathbb{R}$, consider the hypothesis test with rejection region $\{x \in \mathbb{R}^n\colon t(x) \geq c\}$. Let $p\colon \mathbb{R}^n \to [0,1]$ be a function defined by

$$p(x) := \sup_{\theta\in\Theta_0} \mathbf{P}_\theta(t(X) \geq t(x)), \qquad \forall\, x \in \mathbb{R}^n.$$

The ***p*-value** for this set of hypothesis tests is defined to be the statistic $p(X)$.

Remark 5.18. If $c \in \mathbb{R}$ is fixed, then $\beta(\theta) = \mathbf{P}_\theta(X \in C) = \mathbf{P}_\theta(t(X) \geq c)$, by definition of the rejection region C. And the significance level α is defined as

$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \in \Theta_0} \mathbf{P}_\theta(t(X) \geq c).$$

So, $p(x)$ is equal to the significance level of the test where $c = t(x)$. Since α decreases as c increases, we say that $p(x)$ is the smallest significance level such that the hypothesis test rejects the null hypothesis (if α strictly decreases as c increases.)

Remark 5.19. Assume that $Y := t(X)$ is a continuous random variable. Fix $\theta \in \Theta$. For any $c \in \mathbb{R}$, define $F_{-Y}(c) := \mathbf{P}_\theta(-Y \leq c)$. For any $x \in \mathbb{R}^n$ denote $g_\theta(x) := \mathbf{P}_\theta(t(X) \geq t(x)) = \mathbf{P}_\theta(-t(X) \leq -t(x)) = F_{-Y}(-t(x))$. Then $g_\theta(X) = F_{-Y}(-t(X)) = F_{-Y}(-Y)$. So,

$$\mathbf{P}_\theta(g_\theta(X) \leq c) = \mathbf{P}_\theta(F_{-Y}(-Y) \leq c) = \mathbf{P}_\theta(-Y \leq F_{-Y}^{-1}(c)) = F_{-Y}(F_{-Y}^{-1}(c)) = c.$$

So, by definition of $p(x)$, for every $\theta \in \Theta_0$, and for every $c \in [0, 1]$,

$$\mathbf{P}_\theta(p(X) \leq c) \leq \mathbf{P}_\theta(g_\theta(X) \leq c) = c.$$

So, for example, if the null hypothesis is true (i.e. $\theta \in \Theta_0$), then $p(X) \leq .05$ with probability at most .05. If the p-value is observed to be small, then the null hypothesis is believed to be false with high probability. (If the null hypothesis is true, then it is unlikely to observe a small p-value.)

Example 5.20. We continue Example 5.7 and Exercise 2.19. The roulette wheel has 38 spaces and 18 red spaces. Suppose we spin the roulette wheel 5 times resulting in X red outcomes. We model the set of outcomes as a sum of independent $\{0, 1\}$ valued random variables, so that the total number of red outcomes X is a binomial random variable with parameters n, θ with $n = 5$ and $\theta \in [0, 1]$ unknown. Suppose the null hypothesis H_0 is $\{\theta = 1/2\}$, and the alternative hypothesis H_1 is $\{\theta \in [0, 1],\ \theta \neq 1/2\}$.

Consider the hypothesis test with rejection region $C := \{x \in \mathbb{R} \colon x \geq 3\}$. Since Θ_0 consists of a single point, then we define

$$p(x) := \mathbf{P}_{1/2}(X \geq x), \qquad \forall\, x \in \mathbb{R},$$

and the p-value of this test is $p(X)$. So, for example, if we observe that X is 2, i.e. we observe exactly two red outcomes on the roulette wheel, then the reported p-value is

$$\mathbf{P}_{1/2}(X \geq 2) = 1 - \mathbf{P}_{1/2}(X \leq 1) = 1 - (1/2)^5 - 5(1/2)^5 = 1 - 6/32 = .8125.$$

So, in this case, we are not at all confident in rejecting the null hypothesis (we might instead conclude that the null hypothesis is true).

If we observe that X is 4, i.e. we observe exactly four red outcomes on the roulette wheel, then the reported p-value is

$$\mathbf{P}_{1/2}(X \geq 4) = 5(1/2)^5 + (1/2)^5 = 6/32 = .1875.$$

In this case we are more confident in rejecting the null hypothesis.

Exercise 5.21. Suppose X is a binomial distributed random variable with parameters $n = 100$ and $\theta \in [0, 1]$ where θ is unknown. Suppose we want to test the hypothesis H_0 that $\theta = 1/2$ versus the hypothesis H_1 that $\theta \neq 1/2$. Consider the hypothesis test that rejects the null hypothesis if and only if $|X - 50| > 10$.

Using e.g. the central limit theorem, do the following:

- Give an approximation to the significance level α of this hypothesis test
- Plot an approximation of the power function $\beta(\theta)$ as a function of θ.
- Estimate p values for this test when $X = 50$, and also when $X = 70$ or $X = 90$.

5.4. **Generalized Likelihood Ratio Tests.** Let $X_1, \dots, X_n$ be a real-valued random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. We denote the joint distribution of $X_1, \dots, X_n$ as

$$\prod_{i=1}^{n} f_\theta(x_i), \qquad \forall\, 1 \le i \le n.$$

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**. Below we denote $f_\theta(x) = \ell(\theta)$.

The Neyman-Pearson Lemma demonstrates that, when Θ has exactly two points, a likelihood ratio test is UMP among all tests of significance level at most α. When Θ has more than two points, there is an analogue of the likelihood ratio test that has some desirable properties.

Let $\Theta_0 \subseteq \Theta$. When Θ consists of two points $\{\theta_0, \theta_1\}$ and Θ_0 consists of one point θ_0, we defined the likelihood ratio test for the hypothesis H_0 that $\{\theta = \theta_0\}$ in the Neyman-Pearson Lemma 5.9 by its rejection region C'.

$$C' := \{x \in \mathbb{R}^n \colon f_{\theta_1}(x) \ge k f_{\theta_0}(x)\}.$$

Here $k > 0$. Written another way, the rejection region is

$$C' := \{x \in \mathbb{R}^n \colon \sup_{\theta \in \Theta_0^c} f_\theta(x) \ge k \sup_{\theta \in \Theta_0} f_\theta(x)\}.$$

We could use this C' to define a generalized likelihood ratio test, but for technical reasons, the following modification is more convenient.

Definition 5.22 (**Generalized Likelihood Ratio Test**)**.** Let $k \ge 1$. The **generalized likelihood ratio test** of a hypothesis H_0 that $\{\theta \in \Theta_0\}$ is defined by the following rejection region.

$$C := \{x \in \mathbb{R}^n \colon \sup_{\theta \in \Theta} f_\theta(x) \ge k \sup_{\theta \in \Theta_0} f_\theta(x)\}.$$

Intuitively, $\sup_{\theta \in \Theta_0} f_\theta(x)$ chooses the null parameter $\theta \in \Theta_0$ that best fits the data x. So, the generalized likelihood ratio test compares the likelihood of the parameter $\theta \in \Theta$ that best fits the data x, to the likelihood of the null parameter $\theta \in \Theta_0$ that best fits the data x,

Remark 5.23. If $0 < k \le 1$ then $C = \mathbb{R}^n$. That is, all generalized likelihood ratio tests with $0 < k \le 1$ are the same, hence our restriction to $k \ge 1$ in Definition 5.22.

Let D be the set of $x \in \mathbb{R}^n$ such that $\sup_{\theta \in \Theta_0^c} f_\theta(x) \ge \sup_{\theta \in \Theta_0} f_\theta(x)$. If $x \in D$, then $\sup_{\theta \in \Theta} f_\theta(x) = \sup_{\theta \in \Theta_0^c} f_\theta(x)$. So, $C \cap D = C' \cap D$. On D^c, we could have $C \cap D^c \ne C' \cap D^c$. So, at least on the set D, the rejection regions C and C' agree.

Example 5.24. Let $X_1, \dots, X_n$ be a random sample from a Gaussian distribution with known variance $\sigma^2 > 0$ but unknown mean $\mu \in \mathbb{R}$. Fix $\mu_0 \in \mathbb{R}$. Suppose we want to test

the hypothesis H_0 that $\mu = \mu_0$ versus the alternative H_1 that $\mu \neq \mu_0$. That is, $\Theta = \mathbb{R}$, $\Theta_0 = \{\mu_0\}$ and $\Theta_0^c = \Theta_1 = \{\mu \in \mathbb{R} \colon \mu \neq \mu_0\}$. Also, for any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$,

$$f_\mu(x) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

From Example 4.41, the MLE is the sample mean, i.e. for any $x \in \mathbb{R}^n$,

$$\sup_{\mu\in\Theta} f_\mu(x) = f_{\left(\frac{x_1+\cdots+x_n}{n}\right)}(x).$$

Since Θ_0 is just a single point, we can then write the rejection region of the generalized likelihood ratio test as

$$\begin{aligned}
C &:= \{x \in \mathbb{R}^n \colon \sup_{\mu\in\Theta} f_\mu(x) \geq k \sup_{\mu\in\Theta_0} f_\mu(x)\} \\
&= \Big\{x \in \mathbb{R}^n \colon \prod_{i=1}^n e^{-\frac{(x_i - \frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2}{2\sigma^2}} \geq k\Big\} \\
&= \Big\{x \in \mathbb{R}^n \colon e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n \left[(x_i - \frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2\right]} \geq k\Big\} \\
&= \Big\{x \in \mathbb{R}^n \colon \sum_{i=1}^n \Big[(x_i - \frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2\Big] \leq -2\sigma^2 \log k\Big\} \\
&= \Big\{x \in \mathbb{R}^n \colon -n\Big(\frac{1}{n}\sum_{j=1}^n x_j - \mu_0\Big)^2 \leq -2\sigma^2 \log k\Big\} \\
&= \Big\{x \in \mathbb{R}^n \colon \Big|\frac{1}{n}\sum_{j=1}^n x_j - \mu_0\Big| \geq \sqrt{2n^{-1}\sigma^2 \log k}\Big\}.
\end{aligned}$$

So, the test rejects the null hypothesis, unless $\frac{1}{n}\sum_{j=1}^n X_j$ is close to μ_0. As anticipated by Proposition 5.16, the hypothesis test corresponds to confidence intervals for the sample mean. (Above we used the identity $\sum_{i=1}^n \left[(x_i - \frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2\right] = \sum_{i=1}^n \left[(x_i - \mu_0 + \mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2\right] = n(\mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2 - \frac{2}{n}\sum_{i,j=1}(x_i-\mu_0)(x_j-\mu_0) = n(\mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2 - 2n(\frac{1}{n}\sum_{i=1}^n (x_i-\mu_0))^2 = -n(\mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2$.)

Note also that the rejection region of this hypothesis test is a function of a sufficient statistic, since the sample mean is a sufficient statistic for μ by Example 4.11. Intuitively, since the sufficient statistic contains all information about μ, it should not be a surprise that the hypothesis test only needs to check the sufficient statistic.

Denoting $X := (X_1, \ldots, X_n)$, observe that, if H_0 is true, then

$$2\log\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} = \frac{n}{\sigma^2}\Big(\frac{1}{n}\sum_{j=1}^n [X_j - \mu_0]\Big)^2 = \Big(\frac{1}{\sigma\sqrt{n}}\sum_{j=1}^n [X_j - \mu_0]\Big)^2$$

has a chi-squared distribution with one degree of freedom. In fact, this holds asymptotically as $n \to \infty$ in general (see Theorem 5.28 below.)

Finally, note that the p-value for this hypothesis test is

$$p(X), \qquad \text{where} \qquad p(x) := \mathbf{P}_{\theta_0}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} X_j - \mu_0\Big| \geq \Big|\frac{1}{n}\sum_{j=1}^{n} x_j - \mu_0\Big|\Big) \qquad \forall\, x \in \mathbb{R}^n.$$

Exercise 5.25. Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with unknown location parameter $\theta > 0$, i.e. X_1 has density

$$g(x) := 1_{x \geq \theta} e^{-(x-\theta)}, \qquad \forall\, x \in \mathbb{R}.$$

Fix $\theta_0 \in \mathbb{R}$. Suppose we want to test that hypothesis H_0 that $\theta \leq \theta_0$ versus the alternative H_1 that $\theta > \theta_0$. That is, $\Theta = \mathbb{R}$, $\Theta_0 = \{\theta \in \mathbb{R} \colon \theta \leq \theta_0\}$ and $\Theta_0^c = \Theta_1 = \{\theta \in \mathbb{R} \colon \theta > \theta_0$.

- Explicitly describe the rejection region of the generalized likelihood ratio test for this hypothesis. (Hint: it might be easier to describe the region using $x_{(1)} = \min(x_1, \ldots, x_n)$.)
- Prove that $X_{(1)} := \min(X_1, \ldots, X_n)$ is a sufficient statistic for θ.
- (Optional) If H_0 is true, then does

$$2 \log \frac{\sup_{\theta \in \Theta} f_\theta(X_1, \ldots, X_n)}{\sup_{\theta \in \Theta_0} f_\theta(X_1, \ldots, X_n)}$$

converge in distribution to a chi-squared distribution as $n \to \infty$?

Exercise 5.26. Let $X_1, \ldots, X_n$ be a random sample from a Gaussian random variable with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$.

Fix $\mu_0 \in \mathbb{R}$. Suppose we want to test that hypothesis H_0 that $\mu = \mu_0$ versus the alternative H_1 that $\mu \neq \mu_0$.

- Explicitly describe the rejection region of the generalized likelihood ratio test for this hypothesis.
- Give an explicit formula for the p-value of this hypothesis test. (Hint: If S^2 denotes the sample variance and $\overline{X}$ denotes the sample mean, you should then be able to use the statistic $\frac{(\overline{X}-\mu_0)^2}{S^2}$. Since we have an explicit formula for Snedecor's distribution, you should then be able to write an explicit integral formula for the p-value of this test.)

5.5. **Case Study: alpha particle emissions.** The table below demonstrates counts for alpha particle emissions of americium 241. During 1207 disjoint intervals of ten seconds, a number m of alpha particle emission were observed.

m	0, 1 or 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	≥ 17
# of Intervals	18	28	56	105	126	146	164	161	123	101	74	53	23	15	9	5

The number of alpha particle emissions in each of the 1207 intervals is modeled as 1207 i.i.d. Poisson distributed random variables with unknown mean $\lambda > 0$. (So, $\mathbf{P}_\lambda(X = k) = e^{-\lambda}\lambda^k/k!$ for any nonnegative integer $k \geq 0$, and $\lambda > 0$ is unknown.) (There are both mathematical and physical explanations for this assumption which we omit.)

The average number of alpha particles emitted in a ten-second interval of time (averaged over all 1207 intervals) is observed to be 8.392, so we could naively predict that $\lambda \approx 8.392$.

For any integer $k \geq 0$, let $q_k \geq 0$ denote the probability of an alpha particle emission count being k in a ten second time interval, so that $\sum_{k=0}^{\infty} q_k = 1$. And for any $1 \leq j \leq 16$, let p_j be

the probability of the count appearing in the j^{th} column of the table. Then the probability of a count appearing in the $0, 1, 2$ cell in the table is $p_1 := q_0 + q_1 + q_2$, the probability of that count appearing in the 3 cell in the table is $p_2 := q_3$, etc., and the probability of that count appearing in the ≥ 17 cell in the table is $p_{16} = \sum_{j=17}^{\infty} q_j$.

Consider the null hypothesis that $q_k = e^{-\lambda}\lambda^k/k!$ for any $\lambda > 0$, $k \geq 0$, versus the alternative, which includes the assumption that $\sum_{j=1}^{16} p_j = 1$ and $p_j \geq 0$ for all $1 \leq j \leq 16$. Since the table has sixteen entries, we can model the probabilities of the counts by a multinomial distribution, i.e. with 1207 trials of rolling a 16-sided die with unknown probabilities of occurrence of the die rolls. That is, we consider random variables $X_1, \ldots, X_{16}$ defined by the joint distribution

$$f_\theta(x) = f_\theta(x_1, \ldots, x_{16}) := \mathbf{P}(X_1 = x_1, \ldots, X_{16} = x_{16}) = 1207! \prod_{j=1}^{16} \frac{p_j(\theta)^{x_j}}{x_j!},$$

$$\forall\, x_j \in \mathbb{Z},\ x_j \geq 0\ \forall\, 1 \leq j \leq 16,\ \sum_{j=1}^{16} x_j = 1207.$$

To find the supremum of f_θ over all θ, we use Lagrange multipliers with the constraint $\sum_{j=1}^{16} p_j = 1$ and $p_1, \ldots, p_{16} \geq 0$. We have $\frac{\partial f_\theta(x)}{\partial p_j} = \frac{x_j}{p_j} f_\theta(x)$ for all $1 \leq j \leq 16$. Then there exists $\delta \neq 0$ such that $\delta = \frac{\partial f_\theta(x)}{\partial p_j} = \frac{x_j}{p_j} f_\theta(x)$ for all $1 \leq j \leq 16$. That is, at the only interior critical point, we have $x_j = p_j \frac{x_1}{p_1}$ for all $1 \leq j \leq 16$. Summing over j gives $1207 = \frac{x_1}{p_1}$. That is, $p_1 = \frac{x_1}{1207}$. Repeating this argument for any index $1 \leq j \leq 16$ gives

$$p_j = \frac{x_j}{1207}, \qquad \forall\, 1 \leq j \leq 16.$$

Therefore

$$\sup_{\theta \in \Theta} f_\theta(x) = 1207! \prod_{j=1}^{16} \frac{p_j(\theta)^{x_j}}{x_j!} = 1207! \prod_{j=1}^{16} \frac{(x_j/1207)^{x_j}}{x_j!}.$$

(We only found one interior critical point, so we should also argue that this critical point actually is a maximum instead of a minimum. This holds since the likelihood is zero on the boundary of the optimization region, i.e. $f_\theta(x) = 0$ whenever $p_j = 0$ for some $1 \leq j \leq 16$.)

Meanwhile, the supremum over $\theta \in \Theta_0$ can be found by an unconstrained optimization over $\lambda > 0$. (Recall that the first entry of the table has probability $e^{-\lambda}[1 + \lambda + \lambda^2/2]$, and the last entry of the table has probability $e^{-\lambda} \sum_{i=17}^{\infty} \frac{\lambda^k}{k!}$). So,

$$\begin{aligned}
&\sup_{\theta \in \Theta_0} f_\theta(x) \\
&= \sup_{\lambda > 0} 1207! \Big(\prod_{j=2}^{15} \frac{[e^{-\lambda}\lambda^{j+1}/(j+1)!]^{x_j}}{x_j!} \Big) \cdot \frac{(e^{-\lambda}[1 + \lambda + \lambda^2/2])^{x_1}}{x_1!} \cdot \frac{[e^{-\lambda} \sum_{i=17}^{\infty} \frac{\lambda^i}{i!}]^{x_{16}}}{x_{16}!} \\
&= \sup_{\lambda > 0} 1207! \Big(\prod_{j=2}^{15} \frac{[e^{-\lambda}\lambda^{j+1}/(j+1)!]^{x_j}}{x_j!} \Big) \cdot \frac{(e^{-\lambda}[1 + \lambda + \lambda^2/2])^{x_1}}{x_1!} \cdot \frac{[e^{-\lambda}(e^{\lambda} - \sum_{i=0}^{16} \frac{\lambda^i}{i!})]^{x_{16}}}{x_{16}!} \\
&= \sup_{\lambda > 0} 1207! \Big(\prod_{j=2}^{15} \frac{[e^{-\lambda}\lambda^{j+1}/(j+1)!]^{x_j}}{x_j!} \Big) \cdot \frac{(e^{-\lambda}[1 + \lambda + \lambda^2/2])^{x_1}}{x_1!} \cdot \frac{[1 - e^{-\lambda} \sum_{i=0}^{16} \frac{\lambda^i}{i!})]^{x_{16}}}{x_{16}!}.
\end{aligned}$$

Using the data from the above table, with $x_1 = 18, x_2 = 28, \ldots, x_{16} = 5$, we numerically compute the maximum λ to be $\lambda \approx 8.366$, which is very close to the sample mean of 8.392. (Even if you remove the factorials that do not depend on λ, the product of the remaining terms will evaluate to 0 or ∞ on a computer; to fix this issue you can e.g. take the 1/200 power of each product term.) The likelihood ratio is then

$$\frac{\sup_{\theta\in\Theta} f_\theta(x)}{\sup_{\theta\in\Theta_0} f_\theta(x)}$$
$$\approx \Big[\prod_{j=2}^{15}\Big(\frac{x_j/1207}{e^{-8.37}8.37^{j+1}/(j+1)!}\Big)^{x_j}\Big]\Big[\frac{x_1/1207}{[e^{-8.37}(1+8.37+8.37^2/2)]}\Big]^{x_1}\Big[\frac{x_{16}/1207}{[e^{-8.37}\sum_{i=17}^{\infty}\frac{8.37^i}{i!}]}\Big]^{x_{16}}$$

The main question we want to answer is: Is the above Poisson model sensible? That is, does the above Poisson assumption fit the data well? In order to answer this question, we will examine more closely the generalized likelihood ratio. In the case that $X_1, \ldots, X_{16}$ are i.i.d. and $X = (X_1, \ldots, X_{16})$, we know that the quantity

$$2\log\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} \qquad (*)$$

is close to a chi-squared distribution with one degree of freedom, if 16 was replaced by a much larger number. However, $X_1, \ldots, X_{16}$ are not i.i.d., and 16 is not a very large number. Still, 1207 is a fairly large number, so perhaps we can approximately find the distribution of $(*)$ for this reason. Observe

$$2\log\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} = 2\log\prod_{j=1}^{16}\frac{(X_j/1207)^{X_j}}{p_j(\lambda)^{X_j}} = 2\log\prod_{j=1}^{16}\Big(\frac{(X_j/1207)}{p_j(\lambda)}\Big)^{X_j}$$
$$= 2\sum_{j=1}^{16} X_j\log\Big(\frac{X_j/1207}{p_j(\lambda)}\Big) = 2\cdot 1207\sum_{j=1}^{16}\frac{X_j}{1207}\log\Big(\frac{X_j/1207}{p_j(\lambda)}\Big).$$

If H_0 is true, i.e. the data does fit a Poisson distribution, then the MLE for $\theta \in \Theta$ is approximately the same as the MLE for $\lambda > 0$ (i.e. for $\theta \in \Theta_0$), so we have the approximation $X_j/1207 \approx p_j(\lambda)$. So, using the Taylor expansion around $b > 0$ for $h(a) := a\log(a/b)$, we have $h(b) = 0$, $h'(b) = 1$ and $h''(b) = 1/b$, so

$$a\log(a/b) \approx (a-b) + \frac{1}{2b}(a-b)^2.$$

Substituting into the above with $a = X_j/1207$ and $b = p_j(\lambda)$, we get

$$2\log\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} \approx 2\cdot 1207\sum_{j=1}^{16}\Big[\Big(\frac{X_j}{1207} - p_j(\lambda)\Big) + \frac{1}{2}\frac{\Big(\frac{X_j}{1207} - p_j(\lambda)\Big)^2}{p_j(\lambda)}\Big].$$

The first term in the sum is zero since $\sum_{j=1}^{16} X_j = 1207$ and $\sum_{j=1}^{16} p_j(\lambda) = 1$. So,

$$2\log\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} \approx 1207\sum_{j=1}^{16}\frac{\Big(\frac{X_j}{1207} - p_j(\lambda)\Big)^2}{p_j(\lambda)} = \sum_{j=1}^{16}\frac{\Big(X_j - 1207p_j(\lambda)\Big)^2}{1207p_j(\lambda)}.$$

The last quantity is known as **Pearson's chi-squared statistic**. For each $1 \leq j \leq 16$, X_j is a binomial random variable with expected value $1207 p_j(\lambda)$ under H_0. So we can rewrite this statistic as

$$S := \sum_{j=1}^{16} \frac{\Big(X_j - \mathbf{E}_\lambda X_j\Big)^2}{\mathbf{E}_\lambda X_j}.$$

We would like to use this statistic and report its p-value. If S is large, then the data does not follow from a Poisson distribution, so the null hypothesis is false. That is, the test should reject when $S \geq s$ for some $s > 0$. In order to compute the p-value, we will show that the asymptotic distribution of this statistic (as the number of trials $m = 1207$ becomes large) is a chi-squared distribution with $16 - 1 - 1 = 14$ degrees of freedom.

For any given ten-second interval of time, we can record the number of alpha particle emissions as a vector $Y = (Y_1, \ldots, Y_{16})$ of zeros and ones, so $Y_k = 1$ if the count of alpha particles is placed in the k^{th} column of the table, and all other entries of Y are zero. For example, if three emissions are observe then $Y = (0, 1, 0, 0, \ldots, 0)$. We let $M_{ij} := \mathbf{E}(Y_i - \mathbf{E}Y_i)(Y_j - \mathbf{E}Y_j)$ for all $1 \leq i, j \leq 16$ be the covariance matrix of Y. For example, $\mathbf{E}Y_i = p_i$, $\mathbf{E}Y_i^2 = p_i$ and $\mathbf{E}Y_iY_j = 0$ for all $1 \leq i < j \leq m$. We then have

$$M = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & \cdots & -p_1p_{16} \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 & \cdots & -p_2p_{16} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ -p_{16}p_1 & -p_{16}p_2 & -p_{16}p_3 & \cdots & p_{16}(1-p_{16}). \end{pmatrix}$$

This matrix does not have full rank since $\sum_{j=1}^{16} p_j = 1$ implies that M applied to the constant vector is zero. Since this matrix does not have full rank, there will be a technical issue involved in applying the multivariable central limit theorem. So, let us instead examine $Z := (Y_1, \ldots, Y_{15})$. If $p_1, \ldots, p_{16} \neq 0$, the covariance matrix of Z is then

$$R := \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & \cdots & -p_1p_{15} \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 & \cdots & -p_2p_{15} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ -p_{15}p_1 & -p_{15}p_2 & -p_{15}p_3 & \cdots & p_{16}(1-p_{15}) \end{pmatrix}.$$

We can explicitly write an inverse of this matrix, implying that it has full rank:

$$R^{-1} = \begin{pmatrix} p_1^{-1} + p_{16}^{-1} & p_{16}^{-1} & p_{16}^{-1} & \cdots & p_{16}^{-1} \\ p_{16}^{-1} & p_2^{-1} + p_{16}^{-1} & p_{16}^{-1} & \cdots & p_{16}^{-1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ p_{16}^{-1} & p_{16}^{-1} & p_{16}^{-1} & \cdots & p_{15}^{-1} + p_{16}^{-1} \end{pmatrix}.$$

Using again $\sum_{j=1}^{16} X_j = 1207$ and $\sum_{j=1}^{16} p_j(\lambda) = 1$,

$$\begin{aligned} S &= \sum_{j=1}^{16} \frac{\Big(X_j - 1207p_j\Big)^2}{1207p_j} = \sum_{j=1}^{15} \frac{\Big(X_j - 1207p_j\Big)^2}{1207p_j} \qquad + \frac{\Big(X_{16} - 1207p_{16}\Big)^2}{1207p_{16}} \\ &= \sum_{j=1}^{15} \frac{\Big(X_j - 1207p_j\Big)^2}{1207p_j} \qquad + \frac{\Big([p_{16} - 1]1207 - X_j + 1207\Big)^2}{1207p_{16}} \\ &= \sum_{j=1}^{15} \frac{\Big(X_j - 1207p_j\Big)^2}{1207p_j} \qquad + \frac{\Big(\sum_{i=1}^{15}(X_i - 1207p_i)\Big)^2}{1207p_{16}} \\ &= \frac{1}{1207}(X' - 1207p')^T R^{-1}(X' - 1207p'), \end{aligned}$$

where $X' = (X_1, \ldots, X_{15})$ and $p' = (p_1, \ldots, p_{15})$. Letting $Z_1, \ldots, Z_{1207}$ be i.i.d. copies of Z, we have $X_j = \sum_{i=1}^{1207}(Z_i)_j = 1207\overline{Z}_i$. We then have

$$S = 1207(\overline{Z} - p')^T R^{-1}(\overline{Z} - p') = [R^{-1/2}\sqrt{1207}(\overline{Z} - p')]^T R^{-1/2}\sqrt{1207}(\overline{Z} - p').$$

From the multivariable Central Limit Theorem 2.33, $R^{-1/2}\sqrt{1207}(\overline{Z} - p')$ converges to a standard Gaussian random vector, i.e. a vector of 15 i.i.d. standard Gaussian random variables, as $m = 1207$ goes to infinity. It follows that, for fixed $\lambda > 0$, S has the distribution of a chi-squared random variable with 15 degrees of freedom.

In the generalized likelihood ratio, we used λ that is a function of the data $X_1, \ldots, X_{16}$, since we estimated λ using the data. That is, under H_0, we introduce an extra dependence on the random variables $X_1, \ldots, X_{16}$, resulting in one less degree of freedom in the limiting distribution. (For a formal proof of that fact, see A. W. van der Vaart's book, Asymptotic Statistics, Corollary 17.5). So, the distribution of S is approximately a chi-squared random variable with 14 degrees of freedom. From the data, we have

$$\begin{aligned} S = \sum_{j=1}^{16} \frac{\Big(X_j - 1207p_j\Big)^2}{1207p_j} &= \frac{(18 - 1207e^{-8.366}[1 + 8.366 + 8.366^2/2])^2}{1207e^{-8.366}[1 + 8.37 + 8.366^2/2]} \\ &+ \frac{(28 - 1207e^{-8.366}8.366^3/3!)^2}{1207e^{-8.366}8.366^3/3!} + \cdots + \frac{(9 - 1207e^{-8.366}8.366^{16}/16!)^2}{1207e^{-8.366}8.366^{16}/16!} \\ &+ \frac{(5 - 1207[1 - e^{-8.366}\sum_{j=0}^{16} 8.366^j/j!])^2}{1207[1 - e^{-8.366}\sum_{j=0}^{16} 8.366^j/j!]} \end{aligned}$$

We get $S \approx 8.95$. And $\mathbf{P}(S \geq 8.95) \approx .834$. This is a p-value, corresponding to a test that rejects the null hypothesis (that the data follows from a Poisson distribution) when S is large. We therefore accept the null hypothesis, i.e. we believe that the data can be modelled well from the Poisson distribution.

Exercise 5.27. Write down the generalized likelihood ratio estimate for the following alpha particle data, as we did in class for a slightly different data set. The corresponding test treats individual counts of alpha particles as independent Poisson random variables, versus the alternative that the probability of a count appearing in each box of data is a sequence of

nonnegative numbers that sum to one. (In doing so, you should need to compute a maximum likelihood estimate using a computer.)

m	0, 1 or 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	≥ 17
# of Intervals	16	26	58	102	125	146	163	164	120	100	72	54	20	12	10	4

Plot the MLE for the Poisson statistic (i.e. plot the denominator of the generalized likelihood ratio test statistic $\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)}$) as a function of λ.

Finally, compute the value s of Pearson's chi-squared statistic S, and compute the probability that $S \geq s$. Does the probability $\mathbf{P}(S \geq s)$ give you confidence that the null hypothesis is true?

5.6. **Additional Comments.**

Theorem 5.28 (**Limiting Distribution of Generalized Likelihood Ratio Statistic**). *Let $X = (X_1, \ldots, X_n)$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. Fix $\theta_0 \in \Theta \subseteq \mathbb{R}$. Suppose we test the hypothesis H_0 that $\{\theta = \theta_0\}$ versus the alternative $\{\theta \neq \theta_0\}$. Suppose the assumptions of Theorem 4.49 hold. Let $\lambda(X) := \frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)}$ denote the generalized likelihood ratio statistic. If H_0 is true, then $2\log\lambda(X)$ converges in distribution as $n \to \infty$ to a chi-squared random variable with one degree of freedom.*

Proof Sketch. Recall that $\ell(\theta) := \log f_\theta(x)$. Suppose we expand $\ell(\theta)$ in a Taylor series around the random point Y, i.e. assume there exists $h \colon \mathbb{R} \to \mathbb{R}$ such that $\lim_{z\to 0} \frac{h(z)}{z^2} = 0$ and, for all $\theta_0 \in \mathbb{R}$,

$$\ell(\theta_0) = \ell(Y) + \ell'(Y)(\theta_0 - Y) + (1/2)\ell''(Y)(\theta_0 - Y)^2 + h(Y - \theta_0).$$

As in Theorem 4.49, let Y be the MLE. By definition of Y, $\ell'(Y) = 0$. Since $2\log\Lambda(X) = -2\ell(\theta_0) + 2\ell(Y)$, we rearrange the equality to get

$$2\log\Lambda(X) \approx -\ell''(Y)(\theta_0 - Y)^2.$$

As mentioned in Definition 4.24, $\mathbf{E}_{\theta_0}\ell''(\theta_0) = -I_X(\theta_0) = -nI_{X_1}(\theta_0)$. By Theorem 4.48, $Y = Y_n$ converges in probability to the constant θ_0 with respect to $\mathbf{P}_{\theta_0}$ as $n \to \infty$. So, we can approximate $\ell''(Y)$ by $\ell''(\theta_0) \approx -nI_{X_1}(\theta_0)$. That is,

$$2\log\Lambda(X) \approx nI_{X_1}(\theta_0)(\theta_0 - Y)^2.$$

By Theorem 4.49, $\sqrt{n}(Y - \theta_0)$ converges in distribution to a mean zero Gaussian with variance $1/I_{X_1}(\theta_0)$ as $n \to \infty$. Therefore, $2\log\Lambda(X)$ converges in distribution to a chi-squared random variable with one degree of freedom as $n \to \infty$. □

6. Comparing Two Samples

6.1. **Comparing Independent Gaussians.** Suppose $X_1, \ldots, X_n$ is a random sample from a Gaussian random variable X with unknown mean $\mu_X \in \mathbb{R}$ and known variance $\sigma_X^2 > 0$. Suppose $Y_1, \ldots, Y_m$ is a random sample from a Gaussian random variable Y with unknown mean $\mu_Y \in \mathbb{R}$ and known variance $\sigma_Y^2 > 0$.

Assume that $X_1, \dots, X_n$ is independent of $Y_1, \dots, Y_m$, i.e. assume that X, Y are independent. Since X, Y are independent, $X - Y$ is also a Gaussian random variable with mean $\mu_X - \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$, by Example 1.108. Similarly,

$$\frac{\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{m}\sum_{j=1}^m Y_j\right) - \mu_X + \mu_Y}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

is a Gaussian random variable with mean 0 and variance 1. So, for any $t > 0$, we have

$$\mathbf{P}\Big(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{m}\sum_{j=1}^n Y_j - t\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} < \mu_X - \mu_Y < \frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{m}\sum_{j=1}^n Y_j + t\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\Big) = \int_{-t}^t e^{z^2/2}\frac{dz}{\sqrt{2\pi}}.$$

That is, we get confidence intervals for $\mu_X - \mu_Y$, allowing us to obtain estimates on $\mu_X - \mu_Y$.

In the case that the variances are unknown and equal, we can instead integrate Student's t-distribution.

Exercise 6.1. Suppose $X_1, \dots, X_n$ is a random sample from a Gaussian random variable X with unknown mean $\mu_X \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Suppose $Y_1, \dots, Y_m$ is a random sample from a Gaussian random variable Y with unknown mean $\mu_Y \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$.

Assume that $X_1, \dots, X_n$ is independent of $Y_1, \dots, Y_m$, i.e. assume that X, Y are independent.

Assume that $n + m > 2$. Define

$$\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i, \qquad \overline{Y} := \frac{1}{m}\sum_{i=1}^m Y_i,$$

$$S_X^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2, \qquad S_Y^2 := \frac{1}{m-1}\sum_{i=1}^m (Y_i - \overline{Y})^2,$$

$$S^2 := \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

Show that

$$\frac{\overline{X} - \overline{Y} - \mu_X + \mu_Y}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

has Student's t-distribution with $n + m - 2$ degrees of freedom. Deduce the following confidence intervals for the difference of the means

$$\mathbf{P}\Big(\overline{X} - \overline{Y} - tS\sqrt{\frac{1}{n} + \frac{1}{m}} < \mu_X - \mu_Y < \overline{X} - \overline{Y} + tS\sqrt{\frac{1}{n} + \frac{1}{m}}\Big) = \frac{\Gamma(\frac{p+1}{2})}{\sqrt{p}\sqrt{\pi}\Gamma(p/2)}\int_{-t}^t \Big(1 + \frac{s^2}{p}\Big)^{-(p+1)/2} ds,$$

where $p = n + m - 2$.

Exercise 6.2. Suppose you have a random sample of size 6 from a Gaussian random variable with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Suppose this random sample is

$$1,\ 2,\ 3,\ 7,\ 8,\ 9.$$

Explicitly construct a 90% confidence interval for the mean μ.

Then, explicitly construct a 90% confidence interval for the variance $\sigma^2 > 0$.

Your final answer might depend on the function $\Phi(t) := \int_{-\infty}^{t} e^{-x^2/2} dx/\sqrt{2\pi}$, $\Phi\colon \mathbb{R} \to (0,1)$, and/or $\Phi^{-1}\colon (0,1) \to \mathbb{R}$, and/or the corresponding function for Student's t-distribution.

You should not need to use a central limit theorem.

6.2. **Mann-Whitney Test.** A nonparametric method (makes no assumptions about data being sampled from a probability distribution)

Let $n, m > 0$ be integers. Suppose we run some experiment on $m + n$ people e.g. to cure some disease. Among all $m + n$ people, n of them are chosen uniformly at random to be in the control group. The remaining m people are put in a treatment group. The null hypothesis is that the treatment has no effect on people.

Suppose we order the people by integers $1, \ldots, m + n$ so that the integers $1, \ldots, n$ correspond to people in the control group, and integers $n + 1, \ldots, n + m$ correspond to people in the treatment group. Suppose the quality of treatment outcome of person $1 \le i \le n + m$ is some number $x_i \in (0, 1]$ (a high score being a good outcome, and a low score being a bad outcome). Suppose we then linearly order the treatment outcomes as $x_{I_1} \le x_{I_2} \le \cdots \le x_{I_{n+m}}$ where $\{I_1, \ldots, I_{n+m}\} = \{1, \ldots, n + m\}$. For example, the rank of index I_1 is 1, the rank of index I_2 is 2, and in general the rank of index I_j is j. We then use the test statistic

$$Z := \sum_{j=1}^{m+n} j \cdot 1_{\{1 \le I_j \le n\}} = \sum_{I_j=1}^{n} j.$$

That is, Z is the sum of the ranks of people in the control group. Assuming the null hypothesis to be true, all assignments of the ranks $\{1, \ldots, n+m\}$ to the n people $\{1, \ldots, n\}$ in the control group are equally likely. There are $\binom{n+m}{n}$ such assignments, each having probability $\binom{n+m}{n}^{-1}$. For any positive integer k, let $c_{n,m,k}$ be the number of ways that the integer k can be written as a sum of n distinct positive integers among the elements of $\{1, \ldots, n + m\}$ (disregarding the ordering). Then

$$\mathbf{P}(Z = k) = \frac{c_{n,m,k}}{\binom{n+m}{n}}.$$

For example, when $n = 2, m = 3, n + m = 5$, we have $\binom{n+m}{n} = \binom{5}{3} = 10$, $c_{2,3,3} = 1, c_{2,3,4} = 1, c_{2,3,5} = 2, c_{2,3,6} = 2, c_{2,3,7} = 2, c_{2,3,8} = 1, c_{2,3,9} = 1$. So, the distribution of Z can be written explicitly in this case. For general n, m, I am not aware of an elementary formula for $c_{n,m,k}$, though it satisfies the recurrence

$$c_{n,m,k} = c_{n-1,m,k-n} + c_{n,m-1,k-n}, \qquad \forall\, k \ge n \ge 1.$$

Alternatively, $c_{n,m,k}$ can be found by differentiating the following generating function

$$c_{n,m,k} = \frac{1}{n!k!} \frac{\partial^n}{\partial x^n} \frac{\partial^k}{\partial y^k}\Big|_{x=y=0} \prod_{i=1}^{m+n} (1 + xy^i)$$

For example,

$$c_{5,5,20} = 7.$$

Exercise 6.3. Using any method you wish to use, find the number of ways that the integer 30 can be written as a sum of 5 distinct positive integers among the elements of $\{1, 2, \ldots, 10\}$. (Hint: you should probably use a computer.)

Example 6.4. Suppose $n = 2, m = 3, n + m = 5$ and we observe that the Mann-Whitney statistic Z is 7. By definition of Z, $3 \leq Z \leq 9$. The null hypothesis is that the treatment has no effect on people (no "good" effect and no "bad" effect). We should reject the null hypothesis when Z is close to 3 or 9. Consider the hypothesis test that rejects the null hypothesis when $|Z - 6| \geq 2$. We use $|Z - 6|$ as our test statistic in the definition of a p-value. The p-value for the observation that Z is 7 is then

$$\mathbf{P}(|Z - 6| \geq |7 - 6|) = \mathbf{P}(|Z - 6| \geq 1) = 1 - \mathbf{P}(Z = 6) = 1 - \frac{c_{2,3,6}}{\binom{5}{3}} = 1 - \frac{2}{10} = \frac{8}{10} = \frac{4}{5}.$$

So, we are not confident in rejecting the null hypothesis.

On the other hand, if we observe that Z is 4, then the p-value for this observation is

$$\mathbf{P}(|Z - 6| \geq |4 - 6|) = \mathbf{P}(|Z - 6| \geq 2) = \sum_{k \in \{3,4,8,9\}} \frac{c_{2,3,k}}{\binom{5}{3}} = \frac{1}{10}(1 + 1 + 1 + 1) = \frac{4}{10} = \frac{2}{5}.$$

In this case we are a bit more confident in rejection the null hypothesis. As a final example, if we observe that Z is 9, then the p-value for this observation is

$$\mathbf{P}(|Z - 6| \geq |9 - 6|) = \mathbf{P}(|Z - 6| \geq 3) = \sum_{k \in \{3,9\}} \frac{c_{2,3,k}}{\binom{5}{3}} = \frac{1}{10}(1 + 1) = \frac{2}{10} = \frac{1}{5}.$$

So, we are most confident in rejecting the null hypothesis in the case that we observe that Z is 9.

Remark 6.5. Suppose $X_1, \ldots, X_m$ are i.i.d. (treatment outcomes of the m people in the treatment group) and $Y_1, \ldots, Y_n$ are i.i.d. (treatment outcomes of the n people in the control group). Then

$$\sum_{i=1}^{m} \sum_{j=1}^{n} 1_{X_i < Y_j} = \sum_{i=1}^{m} \sum_{j=1}^{n} 1_{X_{(i)} < Y_{(j)}} = \sum_{j=1}^{n} ([\text{rank of } Y_{(j)}] - j) = Z - \frac{n(n+1)}{2}.$$

(The number of $X_{(i)}$ less than $Y_{(j)}$ is equal to the rank of $Y_{(j)}$ minus one, minus the $j - 1$ Y terms ranked below $Y_{(j)}$.) That is,

$$Z = \frac{n(n+1)}{2} + \sum_{i=1}^{m} \sum_{j=1}^{n} 1_{X_i < Y_j}.$$

Under the null hypothesis, $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are i.i.d., and if this distribution is continuous, then $\mathbf{E}1_{X_1 < Y_1} = 1/2$, so

$$\mathbf{E}Z = \frac{n(n+1)}{2} + mn/2 = n\frac{m + n + 1}{2}$$

Also, $\text{Var}1_{X_1<Y_1} = 1/4$, $\mathbf{E}1_{X_1<Y_2}1_{X_1<Y_3} = 1/3$, $\mathbf{E}1_{X_1<Y_2}1_{X_2<Y_4} = 1/4$, so

$$\begin{aligned}
\text{Var}(Z) &= \sum_{i,k=1}^{m}\sum_{j,\ell=1}^{n}\text{Cov}(1_{X_i<Y_j}, 1_{X_k<Y_\ell}) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\text{Cov}(1_{X_i<Y_j}, 1_{X_i<Y_j}) + \sum_{i=1}^{m}\sum_{j\neq\ell}\text{Cov}(1_{X_i<Y_j}, 1_{X_i<Y_\ell}) \\
&\qquad + \sum_{i\neq k}\sum_{j}\text{Cov}(1_{X_i<Y_j}, 1_{X_k<Y_j}) + \sum_{i\neq k}\sum_{j\neq\ell}\text{Cov}(1_{X_i<Y_j}, 1_{X_k<Y_\ell}) \\
&= nm(1/4) + m(n^2-n)(1/12) + (m^2-m)n(1/12) + (m^2-m)(n^2-n)(0) \\
&= nm(1/4 + n/12 - 1/12 + m/12 - 1/12) = nm\frac{1+n+m}{12}.
\end{aligned}$$

So,

$$\frac{Z - \frac{n(m+n+1)}{2}}{\sqrt{nm(m+n+1)/12}}$$

has mean zero and variance one, assuming the null hypothesis is true. And as $m, n \to \infty$, this random variable converges in distribution to a standard Gaussian random variable (though due to lack of independence, this does not follow from the usual Central Limit Theorem.)

Exercise 6.6. Suppose $n = 3, m = 4, n + m = 7$ and we denote Z as the Mann-Whitney statistic. Suppose $Y_1, \ldots, Y_n$ are i.i.d. (treatment outcomes of the n people in the control group) and $X_1, \ldots, X_m$ are i.i.d. (treatment outcomes of the m people in the treatment group). By definition of Z, $1 + 2 + 3 \leq Z \leq 7 + 6 + 5$. The null hypothesis is that the treatment has no effect on people (no "good" effect and no "bad" effect). We should reject the null hypothesis when Z is close to 6 or 18. Consider the (family of) hypothesis tests that rejects the null hypothesis when $|Z - 12| \geq c$ for some constant $c > 0$.

- Compute the p-value for this hypothesis test when $Z = 17$.
- Compute the p-value for this hypothesis test when $Z = 16$.

Now, suppose $n = m = 1000$. Recall that $\mathbf{E}Z = \frac{n(m+n+1)}{2}$. Consider the (family of) hypothesis tests that rejects the null hypothesis when $\left|Z - \frac{n(m+n+1)}{2}\right| \geq c$ for some constant $c > 0$. Using the limiting distribution of Z as an approximation to the distribution of Z itself:

- Approximately compute the p-value for this hypothesis test when $Z = \frac{n(m+n+1)}{2} + 2\sqrt{nm(m+n+1)/12}$.
- Approximately compute the p-value for this hypothesis test when $Z = \frac{n(m+n+1)}{2} + 3\sqrt{nm(m+n+1)/12}$.

6.3. **Comparing Dependent Samples, Signed Rank Test.** Suppose $X_1, \ldots, X_n$ are i.i.d. and $Z_1, \ldots, Z_n$ are i.i.d., but there could be some dependence between these random variables, e.g. X_1 could depend on Z_1. We could imagine data resulting from e.g. a medical study, where X_i is some characteristic of the i^{th} person (e.g. blood pressure) before some treatment is applied, and Z_i is some characteristic of the i^{th} person after a treatment is applied. We would then like to know if the treatment had any significant effect (e.g. by

lowering blood pressure). One way to check for the effect of the treatment is to compute the change in the characteristic $Z_1 - X_1, \ldots, Z_n - X_n$, and then linearly order the absolute values of these differences. That is, we write

$$|Z_{I_1} - X_{I_1}| \le |Z_{I_2} - X_{I_2}| \le \cdots \le |Z_{I_n} - X_{I_n}|,$$

where $\{I_1, \ldots, I_n\} = \{1, \ldots, n\}$. We then use the test statistic

$$W := \sum_{i=1}^{n} \max(i \cdot \text{sign}(Z_{I_i} - X_{I_i}), 0).$$

The null hypothesis is that the treatment has no significant effect. Under the null hypothesis, $Z_1 - X_1$ and $X_1 - Z_1$ have the same distribution. So, $\text{sign}(Z_{I_i} - X_{I_i})$ is equally likely to take the value $+1$ or -1. (For simplicity, just assume that $X_i \neq Z_i$ for all $1 \le i \le n$.) Let us then compute the distribution of W.

Let $Y_1, \ldots, Y_n$ be i.i.d random variables uniformly distributed in $\{-1, 1\}$. Let

$$W_n := \sum_{i=1}^{n} \max(iY_i, 0).$$

We can compute the distribution of W_n recursively, as in Example 1.103. When $n = 1$, $\mathbf{P}(W_1 = 1) = \mathbf{P}(W_1 = 0) = 1/2$. For any $n \ge 1$,

$$\mathbf{P}(W_n = k) = \sum_{j \in \mathbb{Z}} \mathbf{P}(W_{n-1} = j)\mathbf{P}(\max(nY_n, 0) = k - j), \qquad \forall\, k \ge 0.$$

For example, using the Matlab command `conv(,)` to convolve two vectors, we find that

$$\mathbf{P}(W_4 = k) = \begin{cases} 1/8 & \text{, if } k \in \{3, 4, 5, 6, 7\} \\ 1/16 & \text{, if } k \in \{0, 1, 2, 8, 9, 10\}. \end{cases}$$

Exercise 6.7. Let $Y_1, \ldots, Y_n$ be i.i.d random variables uniformly distributed in $\{-1, 1\}$. Let

$$W_n := \sum_{i=1}^{n} \max(iY_i, 0).$$

Explicitly write down the distribution of W_6. (Hint: you should probably use a computer.)

Note that $\mathbf{E}W_n = \sum_{i=1}^{n}(i/2) = \frac{n(n+1)}{4}$ and $\text{Var}(W_n) = \sum_{i=1}^{n}(i^2/4) = \frac{n(n+1)(2n+1)}{24} = n^3/12 + O(n^2)$. So, from the Lindeberg Central Limit Theorem 2.29,

$$\frac{W_n - \frac{n(n+1)}{4}}{\sqrt{n^3/12}}$$

converges to standard Gaussian random variable as $n \to \infty$.

7. Analysis of Variance (ANOVA)

7.1. **General Linear Model.** Let A be an $n \times m$ real matrix of known (deterministic) constants. Let $\beta \in \mathbb{R}^m$ be an unknown vector of (deterministic) constants. And let $\varepsilon \in \mathbb{R}^n$ be a random vector. Our observation of the data is the vector $Y \in \mathbb{R}^n$ defined by

$$Y = A\beta + \varepsilon.$$

The goal is to try to estimate the vector β, when we only have access to Y and A.

Example 7.1 (**One-Way ANOVA**). Let $n_1, n_2, n_3 > 0$ be integers and let $n := n_1+n_2+n_3$. Let $\beta_1, \beta_2, \beta_3 \in \mathbb{R}$ be unknown. Let $\sigma^2 > 0$ be fixed. Let $Y_1, \ldots, Y_n$ be independent random variables such that

- For each $1 \leq i \leq n_1$, Y_i is a Gaussian with mean β_1 and variance σ^2.
- For each $n_1 + 1 \leq i \leq n_1 + n_2$, Y_i is a Gaussian with mean β_2 and variance σ^2.
- For each $n_1 + n_2 + 1 \leq i \leq n$, Y_i is a Gaussian with mean β_3 and variance σ^2.

Then define

$$A := \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

where the matrix A has n_1 rows of the form $(1, 0, 0)$, n_2 rows of the form $(0, 1, 0)$ and n_3 rows of the form $(0, 0, 1)$. Finally, let $\varepsilon \in \mathbb{R}^n$ be a column vector of i.i.d. Gaussians with mean zero and variance σ^2. Then we can write the assumptions of $Y = (Y_1, \ldots, Y_n)$ in matrix form:

$$Y = A\beta + \varepsilon$$

More generally, define

$$A := \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1. \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

where the matrix A has n_j rows with a 1 in the j^{th} entry, for every $1 \leq j \leq p$. Finally, let $\varepsilon \in \mathbb{R}^n$ be a column vector of i.i.d. Gaussians with mean zero and variance σ^2. Then $Y = (Y_1, \ldots, Y_n)$ is a one-way ANOVA of the form

$$Y = A\beta + \varepsilon$$

We can identify two-way ANOVA as a special case of one-way ANOVA. One-way ANOVA considers p groups of data (in the example above, $p = 3$, e.g. red birds, blue birds and green birds). Two-way ANOVA also considers groups of data but sorted according to two characteristics, e.g. red large birds, red small birds, blue large birds, blue small birds, etc.)

Example 7.2 (**Linear Regression**). Let $\beta_1, \beta_2 \in \mathbb{R}$ be unknown. Let $x_1, \ldots, x_n \in \mathbb{R}$ be fixed constants. Let $\sigma^2 > 0$ be fixed. Then define

$$A := \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Finally, let $\varepsilon \in \mathbb{R}^n$ be a column vector of i.i.d. Gaussians with mean zero and variance σ^2. Then the equation

$$Y = A\beta + \varepsilon$$

can be written as

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \qquad \forall\, 1 \le i \le n.$$

That is, x_i and Y_i are observed for all $1 \le i \le n$, and there is an (unknown) linear relationship between these data.

More generally, Let $\beta_0, \ldots, \beta_p \in \mathbb{R}$ be unknown. Let $\{x_{ij}\}_{1 \le i \le n, 1 \le j \le p} \in \mathbb{R}$ be fixed constants. Let $\sigma^2 > 0$ be fixed. Then define

$$A := \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

Finally, let $\varepsilon \in \mathbb{R}^n$ be a column vector of i.i.d. Gaussians with mean zero and variance σ^2. Then the equation

$$Y = A\beta + \varepsilon$$

can be written as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad \forall\, 1 \le i \le n.$$

That is, $\{x_{ij}\}$ and Y_i are observed for all i, j, and there is an (unknown) linear relationship between these data.

7.2. **One-Way ANOVA Hypothesis Testing.** For any $1 \le j \le p$, denote $m_j := n_1 + \cdots + n_j$. In One-Way ANOVA, we have unknown constants $\beta_1, \ldots, \beta_p$ that we would like to find, we have i.i.d. Gaussians $\varepsilon_1, \ldots, \varepsilon_{m_p}$ with mean zero and variance $\sigma^2 > 0$ and we observe $Y_1, \ldots, Y_{m_p}$ where

$$\begin{aligned} Y_i &= \beta_1 + \varepsilon_i, && \forall\, 1 \le i \le m_1 \\ Y_i &= \beta_2 + \varepsilon_i, && \forall\, m_1 + 1 \le i \le m_2 \\ &\vdots \\ Y_i &= \beta_p + \varepsilon_i, && \forall\, m_{p-1} + 1 \le i \le m_p \end{aligned}$$

$$\overline{Y}_j := \frac{1}{n_j} \sum_{i=m_{j-1}+1}^{m_j} Y_i.$$

That is, $\overline{Y}_j$ is the sample mean of the random variables that each have mean β_j. So,

$$\mathbf{E}\overline{Y}_j = \beta_j, \qquad \forall\, 1 \le j \le p.$$

We know from Section 6.1 that, for any $1 \leq j < k \leq p$,

$$\frac{\overline{Y}_j - \overline{Y}_k - (\beta_j - \beta_k)}{\sigma\sqrt{\frac{1}{n_j} + \frac{1}{n_k}}}$$

is a standard Gaussian random variable, so we can get confidence intervals for $\beta_j - \beta_k$ from this fact. More generally, for any constants $c_1, \ldots, c_p$ that are not all zero,

$$\frac{\sum_{j=1}^p c_j \overline{Y}_j - \sum_{j=1}^p c_j \beta_j}{\sigma\sqrt{\sum_{j=1}^p \frac{c_j^2}{n_j}}}$$

is a standard Gaussian random variable, so we can get confidence intervals for $\sum_{j=1}^p c_j \beta_j$ from this fact.

For any $1 \leq j \leq p$, denote the j^{th} sample variance as

$$S_j^2 := \frac{1}{n_j - 1} \sum_{i=m_{j-1}+1}^{m_j} (Y_i - \overline{Y}_j)^2.$$

Recall also from Exercise 6.1 that, for any $1 \leq j < k \leq p$,

$$\frac{\overline{Y}_j - \overline{Y}_k - (\beta_j - \beta_k)}{S\sqrt{\frac{1}{n_j} + \frac{1}{n_k}}}$$

has Student's t-distribution with $n_j + n_k - 2$ degrees of freedom, where

$$S^2 := \frac{(n_j - 1)S_j^2 + (n_k - 1)S_k^2}{n_j + n_k - 2}.$$

More generally, for any constants $c_1, \ldots, c_p$ that are not all zero,

$$\frac{\sum_{j=1}^p c_j \overline{Y}_j - \sum_{j=1}^p c_j \beta_j}{S\sqrt{\sum_{j=1}^p \frac{c_j^2}{n_j}}} \qquad (*)$$

has Student's t-distribution with $\left(\sum_{j=1}^p n_j\right) - p = m_p - p$ degrees of freedom, where

$$S^2 := \frac{\sum_{j=1}^p (n_j - 1)S_j^2}{m_p - p}.$$

Now, suppose we want to test the hypothesis that $\beta_1 = \cdots = \beta_p$, versus the alternative. We then can consider the statistic $(*)$ for any $c_1, \ldots, c_p$ with $\sum_{i=1}^p c_i = 0$, as the following lemma shows.

Lemma 7.3. *The following two conditions are equivalent.*

- $\beta_1 = \cdots = \beta_p$
- *For any $c_1, \ldots, c_p \in \mathbb{R}$ with $\sum_{i=1}^p c_i = 0$, we have*

$$\sum_{i=1}^p c_i \beta_i = 0.$$

Proof. If the first condition holds, then $\sum_{i=1}^p c_i\beta_i = \beta_1 \sum_{i=1}^p c_i = 0$.

If the second condition holds, then fix any $1 \le i < j \le p$, and set $c_i = 1$, $c_j = -1$ and $c_k = 0$ for all other $k \in \{1, \dots, p\}$. The second condition says $\beta_i - \beta_j = 0$, i.e. $\beta_i = \beta_j$, i.e. the first condition holds. □

The null hypothesis that $\beta_1 = \cdots = \beta_p$ is then equivalent to: For any $c_1, \dots, c_p \in \mathbb{R}$ with $\sum_{i=1}^p c_i = 0$, we have

$$\sum_{i=1}^p c_i\beta_i = 0.$$

Proposition 7.4. *Define*

$$F := \sup_{c_1,\dots,c_p \in \mathbb{R}:\ \sum_{i=1}^p c_i = 0} \frac{\left(\sum_{j=1}^p c_j \overline{Y}_j - \sum_{j=1}^p c_j\beta_j\right)^2}{S^2 \sum_{j=1}^p \frac{c_j^2}{n_j}}$$

Then

$$F = \frac{1}{S^2} \sum_{j=1}^p n_j((\overline{Y_j} - \overline{Y}) - (\beta_j - \overline{\beta}))^2,$$

where $\overline{Y} = \frac{1}{m_p}\sum_{i=1}^{m_p} Y_i$, *and* $\overline{\beta} = \mathbf{E}\overline{Y} = \frac{1}{m_p}\sum_{i=1}^{m_p} \mathbf{E}Y_i = \frac{1}{m_p}\sum_{j=1}^p n_j\beta_j$.

Moreover, $F/(p-1)$ *has Snedecor's f-distribution with* $p-1$ *and* $m_p - p$ *degrees of freedom. (For a definition of this distribution, see Exercise 3.11.)*

Proof. Apply Lemma 7.6 with $a_i = n_i^{-1}$, $b_i := \overline{Y}_i - \beta_i \ \forall\ 1 \le i \le p$, noting that

$$\frac{\left(\sum_{j=1}^p c_j \overline{Y}_j - \sum_{j=1}^p c_j\beta_j\right)^2}{\sum_{j=1}^p \frac{c_j^2}{n_j}} = \frac{t^2}{\sum_{i=1}^p a_i c_i^2} = \sum_{\ell=1}^p a_\ell^{-1}\left(b_\ell - \frac{\sum_{j=1}^p b_j a_j^{-1}}{\sum_{k=1}^p a_k^{-1}}\right)^2$$

$$= \sum_{\ell=1}^p n_\ell\left(b_\ell - \frac{\sum_{j=1}^p b_j n_j}{\sum_{k=1}^p n_k}\right)^2 = \sum_{j=1}^p n_j((\overline{Y_j} - \overline{Y}) - (\beta_i - \overline{\beta}))^2$$

Finally, (a generalization of) Proposition 3.7 implies that the numerator and denominator of F are independent, and Exercise 3.11 completes the proof. □

Remark 7.5. Under the null hypothesis that $\beta_1 = \cdots = \beta_p$, we have $\beta_1 = \cdots = \beta_p = \overline{\beta}$, so that

$$F = \frac{1}{S^2} \sum_{j=1}^p n_j(\overline{Y_j} - \overline{Y})^2,$$

That is, F is now a statistic (since it no longer depends on any unknown parameters).

Lemma 7.6. *Let* $a_1, \dots, a_n > 0$, *let* $b_1, \dots, b_n \in \mathbb{R}$ *and let* $t \neq 0$. *Suppose we minimize*

$$\frac{1}{2}\sum_{i=1}^n a_i c_i^2$$

subject to the constraints

$$\sum_{i=1}^n c_i = 0, \qquad \sum_{i=1}^n c_i b_i = t.$$

Then the minimum value of this problem occurs when

$$\sum_{i=1}^{n} a_i c_i^2 = \frac{t^2}{\sum_{i=1}^{n} a_i^{-1}\Big(b_i - \frac{\sum_{j=1}^{n} b_j a_j^{-1}}{\sum_{k=1}^{n} a_k^{-1}}\Big)^2}.$$

$$c_i = \frac{t a_i^{-1}\Big(b_i - \frac{\sum_{j=1}^{n} b_j a_j^{-1}}{\sum_{k=1}^{n} a_k^{-1}}\Big)}{\sum_{\ell=1}^{n} a_\ell^{-1}\Big(b_\ell - \frac{\sum_{j=1}^{n} b_j a_j^{-1}}{\sum_{k=1}^{n} a_k^{-1}}\Big)^2}, \qquad \forall\, 1 \le i \le n.$$

Proof. By Lagrange multipliers, there exists $\lambda_1, \lambda_2 \in \mathbb{R}$ such that

$$a_i c_i = \lambda_1 + \lambda_2 b_i, \qquad \forall\, 1 \le i \le n.$$

Dividing by a_i, summing over i and using the constraints we obtain

$$0 = \lambda_1 \sum_{i=1}^{n} a_i^{-1} + \lambda_2 \sum_{i=1}^{n} b_i a_i^{-1}, \qquad \lambda_1 = -\lambda_2 \frac{\sum_{i=1}^{n} b_i a_i^{-1}}{\sum_{i=1}^{n} a_i^{-1}}$$

Multiplying by c_i and summing over i,

$$\sum_{i=1}^{n} a_i c_i^2 = \lambda_2 t.$$

So,

$$c_i = \frac{1}{a_i}(\lambda_1 + \lambda_2 b_i) = \frac{1}{a_i}\lambda_2\Big(-\frac{\sum_{j=1}^{n} b_j a_j^{-1}}{\sum_{k=1}^{n} a_k^{-1}} + b_i\Big) = \frac{\sum_{\ell=1}^{n} a_\ell c_\ell^2}{t a_i}\Big(-\frac{\sum_{j=1}^{n} b_j a_j^{-1}}{\sum_{k=1}^{n} a_k^{-1}} + b_i\Big).$$

Squaring, multiplying by a_i and summing over i,

$$\sum_{i=1}^{n} a_i c_i^2 = \frac{1}{t^2}\Big(\sum_{k=1}^{n} a_k c_k^2\Big)^2 \sum_{i=1}^{n} a_i^{-1}\Big(-\frac{\sum_{j=1}^{n} b_j a_j^{-1}}{\sum_{k=1}^{n} a_k^{-1}} + b_i\Big)^2.$$

That is,

$$\sum_{i=1}^{n} a_i c_i^2 = \frac{t^2}{\sum_{i=1}^{n} a_i^{-1}\Big(b_i - \frac{\sum_{j=1}^{n} b_j a_j^{-1}}{\sum_{k=1}^{n} a_k^{-1}}\Big)^2}.$$

Finally,

$$c_i = \frac{t a_i^{-1}\Big(b_i - \frac{\sum_{j=1}^{n} b_j a_j^{-1}}{\sum_{k=1}^{n} a_k^{-1}}\Big)}{\sum_{\ell=1}^{n} a_\ell^{-1}\Big(b_\ell - \frac{\sum_{j=1}^{n} b_j a_j^{-1}}{\sum_{k=1}^{n} a_k^{-1}}\Big)^2}, \qquad \forall\, 1 \le i \le n.$$

Since we are minimizing a convex function subject to a linear constraint, and we found one critical point, this critical point must be the unique global minimum, by (a modification of) Exercise 4.37. □

7.3. **Linear Regression.**

Exercise 7.7. In statistics and other applications, we can be presented with data points $(x_1, y_1), \ldots, (x_n, y_n)$. We would like to find the line $y = mx + b$ which lies "closest" to all of these data points. Such a line is known as a **linear regression**. There are many ways to define the "closest" such line. The standard method is to use **least squares minimization**. A line which lies close to all of the data points should make the quantities $(y_i - mx_i - b)$ all very small. We would like to find numbers m, b such that the following quantity is minimized:

$$f(m, b) = \sum_{i=1}^{n} (y_i - mx_i - b)^2.$$

Using the second derivative test, show that the minimum value of f is achieved when

$$m = \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{j=1}^n y_j\right) - n\left(\sum_{k=1}^n x_k y_k\right)}{\left(\sum_{i=1}^n x_i\right)^2 - n\left(\sum_{j=1}^n x_j^2\right)} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{j=1}^n (x_j - \overline{x})^2}.$$

$$b = \frac{1}{n}\left(\sum_{i=1}^n y_i - m\sum_{j=1}^n x_j\right) = \overline{y} - m\overline{x}.$$

Briefly explain why this is actually the minimum value of $f(m, b)$. (You are allowed to use the inequality $(\sum_{i=1}^n x_i)^2 \leq n(\sum_{i=1}^n x_i^2)$.)

From Example 7.2, we originally presented linear regression as the following problem. Let $\beta_1, \beta_2 \in \mathbb{R}$ be unknown. Let $x_1, \ldots, x_n \in \mathbb{R}$ be fixed constants. Let $\sigma^2 > 0$ be fixed. Let $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. Gaussians with mean zero and variance σ^2. Then suppose we observe

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \qquad \forall\, 1 \leq i \leq n.$$

That is, x_i and Y_i are observed for all $1 \leq i \leq n$, and there is an (unknown) linear relationship between these data.

The task is to estimate β_1, β_2. Suppose we restrict only to linear estimators, i.e. estimators of the form

$$\sum_{i=1}^n c_i Y_i$$

where $c_1, \ldots, c_n \in \mathbb{R}$, and we try to find unbiased linear estimator of the smallest variance (similar to a UMVU, but restricted to linear estimators).

Theorem 7.8. *Let $c_1, \ldots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i Y_i$ is an unbiased estimator of β_2. Suppose*

$$\mathrm{Var}(\sum_{i=1}^n c_i Y_i) \leq \mathrm{Var}(\sum_{i=1}^n c_i' Y_i),$$

for all $c_1', \ldots, c_n' \in \mathbb{R}$. Then

$$\sum_{i=1}^n c_i Y_i = \frac{\sum_{i=1}^n (Y_i - \frac{1}{n}\sum_{j=1}^n Y_j)(x_i - \frac{1}{n}\sum_{j=1}^n x_j)}{\sum_{k=1}^n (x_k - \frac{1}{n}\sum_{\ell=1}^n x_\ell)^2}$$

Let $c_1, \dots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i Y_i$ is an unbiased estimator of β_1. Suppose

$$\mathrm{Var}(\sum_{i=1}^n c_i Y_i) \le \mathrm{Var}(\sum_{i=1}^n c_i' Y_i),$$

for all $c_1', \dots, c_n' \in \mathbb{R}$. Then

$$\sum_{i=1}^n c_i Y_i = \frac{1}{n}\sum_{i=1}^n Y_i - \frac{\sum_{i=1}^n (Y_i - \sum_{j=1}^n Y_j)(x_i - \frac{1}{n}\sum_{j=1}^n x_j)}{\sum_{k=1}^n (x_k - \frac{1}{n}\sum_{\ell=1}^n x_\ell)^2} \cdot \frac{1}{n}\sum_{i=1}^n x_i.$$

Proof. **Step 1**. Since

$$\mathbf{E}\sum_{i=1}^n c_i Y_i = \sum_{i=1}^n c_i(\beta_1 + \beta_2 x_i),$$

an unbiased linear estimator of β_2 satisfies

$$\sum_{i=1}^n c_i = 0, \qquad \sum_{i=1}^n c_i x_i = 1. \qquad (*)$$

and the variance of this estimator is

$$\mathrm{Var}\sum_{i=1}^n c_i Y_i = \sum_{i=1}^n c_i^2 \mathrm{Var} Y_i = \sigma^2 \sum_{i=1}^n c_i^2.$$

Suppose we minimize this quantity subject to the constraint $(*)$. Lemma 7.6 with $t = 1$, $b_i = x_i$ and $a_i = 1$ for all i implies that this minimum occurs when

$$c_i = \frac{t a_i^{-1}\left(b_i - \frac{\sum_{j=1}^n b_j a_j^{-1}}{\sum_{k=1}^n a_k^{-1}}\right)}{\sum_{\ell=1}^n a_\ell^{-1}\left(b_\ell - \frac{\sum_{j=1}^n b_j a_j^{-1}}{\sum_{k=1}^n a_k^{-1}}\right)^2} = \frac{x_i - \frac{1}{n}\sum_{j=1}^n x_j}{\sum_{\ell=1}^n \left(x_\ell - \frac{1}{n}\sum_{j=1}^n x_j\right)^2}, \qquad \forall\, 1 \le i \le n.$$

$$\sum_{i=1}^n c_i Y_i = \frac{\sum_{i=1}^n Y_i (x_i - \frac{1}{n}\sum_{j=1}^n x_j)}{\sum_{k=1}^n (x_k - \frac{1}{n}\sum_{\ell=1}^n x_\ell)^2} = \frac{\sum_{i=1}^n (Y_i - \sum_{j=1}^n Y_j)(x_i - \frac{1}{n}\sum_{j=1}^n x_j)}{\sum_{k=1}^n (x_k - \frac{1}{n}\sum_{\ell=1}^n x_\ell)^2}.$$

Step 2. Since

$$\mathbf{E}\sum_{i=1}^n c_i Y_i = \sum_{i=1}^n c_i(\beta_1 + \beta_2 x_i),$$

an unbiased linear estimator of β_1 satisfies

$$\sum_{i=1}^n c_i = 1, \qquad \sum_{i=1}^n c_i x_i = 0. \qquad (**)$$

and the variance of this estimator is

$$\mathrm{Var}\sum_{i=1}^n c_i Y_i = \sum_{i=1}^n c_i^2 \mathrm{Var} Y_i = \sigma^2 \sum_{i=1}^n c_i^2.$$

Suppose we minimize this quantity subject to the constraint $(**)$. Lemma 7.6 with variables $c_i' = c_i x_i$, $b_i = 1/x_i$ $a_i = 1/x_i^2$, with $t = 1$, for all i implies that this minimum occurs when

$$c_i' = \frac{t a_i^{-1}\left(b_i - \frac{\sum_{j=1}^n b_j a_j^{-1}}{\sum_{k=1}^n a_k^{-1}}\right)}{\sum_{\ell=1}^n a_\ell^{-1}\left(b_\ell - \frac{\sum_{j=1}^n b_j a_j^{-1}}{\sum_{k=1}^n a_k^{-1}}\right)^2} = \frac{x_i^2\left(x_i^{-1} - \frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)}{\sum_{\ell=1}^n x_\ell^2\left(x_\ell^{-1} - \frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)^2} = \frac{x_i\left(1 - x_i\frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)}{\sum_{\ell=1}^n \left(1 - x_\ell\frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)^2}.$$

$$c_i = \frac{\left(1 - x_i\frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)}{\sum_{\ell=1}^n \left(1 - x_\ell\frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)^2} = \frac{\left(1 - x_i\frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)}{\sum_{\ell=1}^n \left(1 - 2x_\ell\frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2} + x_\ell^2\left(\frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)^2\right)} = \frac{\left(1 - x_i\frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)}{n - \frac{\left(\sum_{j=1}^n x_j\right)^2}{\sum_{k=1}^n x_k^2}}.$$

$$\begin{aligned}
\sum_{i=1}^n c_i Y_i &= \frac{\sum_{i=1}^n Y_i - \sum_{i=1}^n Y_i x_i \frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}}{n - \frac{\left(\sum_{j=1}^n x_j\right)^2}{\sum_{k=1}^n x_k^2}} = \frac{\sum_{i=1}^n Y_i \sum_{k=1}^n x_k^2 - \sum_{i=1}^n Y_i x_i \sum_{j=1}^n x_j}{n\sum_{k=1}^n x_k^2 - (\sum_{j=1}^n x_j)^2} \\
&= \frac{\sum_{i=1}^n Y_i(\sum_{k=1}^n x_k^2 - \frac{1}{n}(\sum_{j=1}^n x_j)^2) + \sum_{i=1}^n Y_i \frac{1}{n}(\sum_{j=1}^n x_j)^2 - \sum_{i=1}^n Y_i x_i \sum_{j=1}^n x_j}{n\sum_{k=1}^n x_k^2 - (\sum_{j=1}^n x_j)^2} \\
&= \frac{1}{n}\sum_{i=1}^n Y_i + \frac{1}{n}\sum_{\ell=1}^n x_\ell \frac{\sum_{i=1}^n Y_i \sum_{j=1}^n x_j - n\sum_{i=1}^n Y_i x_i}{n\sum_{k=1}^n x_k^2 - (\sum_{j=1}^n x_j)^2}
\end{aligned}$$

□

7.4. **Logistic Regression.** Denote the **logistic function** as

$$h(x) := \frac{1}{1 + e^{-x}}, \qquad \forall\, x \in \mathbb{R}.$$

Note that $\lim_{x\to\infty} h(x) = 1$ and $\lim_{x\to-\infty} h(x) = 0$.

Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables. Let $g\colon \mathbb{R} \to \{0, 1\}$ be an unknown function, and let $Y_i := g(X_i)$ for all $1 \le i \le n$. For example, $X_1, \ldots, X_n$ could be the blood pressures of n people, and $g(X_i) = 1$ if person $i \in \{1, \ldots, n\}$ has had a heart attack, whereas $g(X_i) = 0$ if person i has not had a heart attack. In this way, g classifies the data has having or not having a certain trait. For another example, X_i could be some characteristic of the i^{th} received email, $g(X_i) = 1$ if email $i \in \{1, \ldots, n\}$ is spam, whereas $g(X_i) = 0$ if email i is not spam.

By our assumptions, $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli random variables with some unknown probability $0 \le p \le 1$ such that $p = \mathbf{P}(Y_1 = 1)$. Since the logistic function smoothly transitions from value 0 to value 1, we make the heuristic assumption that there are some unknown parameters $a, b \in \mathbb{R}$ such that

$$p \approx h(ax + b) \approx g(x).$$

The likelihood function is then

$$\ell(a, b) := \prod_{i=1}^n p^{y_i}(1-p)^{1-y_i} = \prod_{i=1}^n [h(ax_i + b)]^{y_i}[1 - h(ax_i + b)]^{1-y_i},$$
$$\forall\, x_1, \ldots, x_n \in \mathbb{R}, \quad \forall\, y_1, \ldots, y_n \in \{0, 1\}.$$

From Exercise 7.9, the log-likelihood function has at most one global maximum. So, if the MLE exists, it is unique.

Exercise 7.9. Let

$$h(x) := \frac{1}{1+e^{-x}}, \qquad \forall\, x \in \mathbb{R}.$$

Fix $x \in \mathbb{R}$ and $y \in [0,1]$. Define $t\colon \mathbb{R}^2 \to \mathbb{R}$ by

$$t(a,b) := \log\Big([h(ax+b)]^y[1-h(ax+b)]^{1-y}\Big), \qquad \forall\, a,b \in \mathbb{R}.$$

Show that t is concave. Conclude that t has at most one global maximum.

Exercise 7.10. 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

$$t = s(u).$$

8. Appendix: Results from Analysis

Theorem 8.1. ***(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

$$\big|f^{(k)}(x)\big| \le \frac{c_{jk}}{1+|x|^j}, \qquad \forall\, x \in \mathbb{R}.$$

Proposition 8.2 (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):

$$\begin{aligned}
\|\phi_\varepsilon * f - cf\|_p &= \left\| \int_{\mathbb{R}} \phi_\varepsilon(y)(f(x-y) - f(x))dy \right\|_{p,dx} \\
&\le \int_{\mathbb{R}} |\phi_\varepsilon(y)|\,\|f(x-y) - f(x)\|_{p,dx}\,dy \quad \text{, by Theorem. 8.1} \\
&= \int_{\mathbb{R}} |\phi(y)|\,\|f(x-\varepsilon y) - f(x)\|_{p,dx}\,dy, \text{ changing variables.}
\end{aligned}$$

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 2.40, 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,

$$\int_{\mathbb{R}} |\phi(y)|\,|f(x-\varepsilon y)-f(x)|\,dy = \int_{\{y\in\mathbb{R}:\,|y|>m\}} (\cdots) + \int_{\{y\in\mathbb{R}:\,|y|\le m\}} (\cdots)$$
$$\le 2\,\|f\|_\infty \int_{\{y\in\mathbb{R}:\,|y|>m\}} |\phi(y)|\,dy + \int_{\{y\in\mathbb{R}:\,|y|\le m\}} |\phi(y)|\,\frac{\eta}{\|\phi\|_1} \le \eta+\eta = 2\eta.$$

Proof of (e): Let $h > 0$ and $x \in \mathbb{R}$. Then

$$\left|\frac{(f*g)(x+h)-(f*g)(x)}{h} - (f'*g)(x)\right| \le \left\|\frac{f(x+h)-f(x)}{h} - f'(x)\right\|_{\infty,dx} \|g\|_1$$
$$\le \left\|\frac{1}{h}\int_x^{x+h} (x+h-t)f''(t)dt\right\|_{\infty,dx} \|g\|_1 \le |h|\,\|f''\|_\infty\,\|g\|_1\,.$$

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 8.3 (**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

$$\int_{\mathbb{R}} e^{ix\xi}\left(\int_{\mathbb{R}} f(x-y)g(y)dy\right)dx = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{ix\xi}f(x-y)dxg(y)dy$$
$$\stackrel{(b)}{=} \int_{\mathbb{R}} e^{i\xi y}\widehat{f}(\xi)g(y)dy = \widehat{f}(\xi)\int_{\mathbb{R}} e^{i\xi y}g(y)dy = \widehat{f}(\xi)\widehat{g}(\xi).$$

Proof of (e): Let $h > 0$. Using part (b) and the Dominated Convergence Theorem 2.40,

$$\frac{\widehat{f}(\xi+h)-\widehat{f}(\xi)}{h} \stackrel{(b)}{=} \mathcal{F}\left[\left(\frac{e^{ixh}-1}{h}\right)f(x)\right](\xi) \to \mathcal{F}[ixf(x)](\xi)\,, \text{as } h\to 0.$$

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 8.4. *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 8.3(a),(d) and Proposition 8.2(a),

$$\left|\widehat{\phi_\varepsilon}(\xi)\widehat{f}(\xi) - \widehat{f}(\xi)\right| = \left|\widehat{\phi_\varepsilon * f}(\xi) - \widehat{f}(\xi)\right| \le \int_{\mathbb{R}} |\phi_\varepsilon * f(x) - f(x)|\,dx \to 0,$$

as $\varepsilon \to 0$. Combining this statement with Proposition 8.3(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 8.3(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 8.3 shows that $\widehat{f}$ is k times differentiable for any $k \ge 1$, since f is a Schwartz function. And part (b) of the current Proposition says that $f^{(k)}$ vanishes at infinity for any $k \ge 1$, so repeated application of Proposition 8.3(f) shows that $\widehat{f}$ is a Schwartz function. □

Exercise 8.5. 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 8.6 (**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 8.3(c) and Proposition 8.4(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 8.3(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 8.3(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 2.40. And the right side tends to f uniformly in x by Proposition 8.2(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 8.7 (**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 $(*)$. □

Proposition 8.8 (**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 $|\partial f(\theta, x)/\partial\theta| \le g(x)$ 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 $\int_0^\theta \int_{\mathbb{R}^n} |h(t, x)|\, dxdt < \infty$ for any $\theta \in \mathbb{R}$. By Fubini's Theorem 1.79,

$$\int_0^\theta \int_{\mathbb{R}^n} h(t, x)dxdt = \int_{\mathbb{R}^n} \int_0^\theta h(t, x)dtdx = \int_{\mathbb{R}^n} h_0(\theta, x)dx < \infty.$$

Taking derivatives in θ of both sides and applying Lebesgue's Fundamental Theorem of Calculus, Theorem 1.42 (twice) concludes the proof. □

9. Appendix: Convergence in Distribution, Characteristic Functions

Definition 9.1 (**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 9.2. *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,

$$\left| \mathbf{E}g(X) - \sum_{i=1}^{m-1} c_i \Big(F_X(t_{i+1}) - F_X(t_i) \Big) \right| = |\mathbf{E}g(X) - \mathbf{E}g_\varepsilon(X)| \leq \mathbf{E}\,|g(X) - g_\varepsilon(X)| \leq \varepsilon.$$

The same holds replacing X with any of $X_1, X_2, \ldots$. So, applying the triangle inequality,

$$\begin{aligned}
&\limsup_{n\to\infty} |\mathbf{E}g(X_n) - \mathbf{E}g(X)| \\
&\quad \leq \limsup_{n\to\infty} |\mathbf{E}g(X_n) - \mathbf{E}g_\varepsilon(X_n)| + |\mathbf{E}g_\varepsilon(X_n) - \mathbf{E}g_\varepsilon(X)| + |\mathbf{E}g_\varepsilon(X) - \mathbf{E}g(X)| \\
&\quad \leq 2\varepsilon + \limsup_{n\to\infty} \sum_{i=1}^{m-1} |c_i|\, |F_{X_n}(t_{i+1}) - F_X(t_{i+1}) - [F_{X_n}(t_i) - F_X(t_i)]| = 2\varepsilon.
\end{aligned}$$

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 9.3. *Let $\mu_1, \mu_2, \dots$ 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 $\mu_1, \mu_2, \dots$ is **tight**: $\forall\ \varepsilon > 0$, $\exists\ m = m(\varepsilon) > 0$ such that*

$$\limsup_{n\to\infty} (1 - \mu_n([-m, m])) \le \varepsilon.$$

Exercise 9.4. 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 9.3.

(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 $n \ge 1$, and such that $Z_1, Z_2, \dots$ converges almost surely to Z. (Hint: use Exercise 3.19.)

(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 $\liminf_{n\to\infty} \mathbf{E}g(X_n) \ge \mathbf{E}g(X)$.

(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 $|X_n| \le Y$ for all $n \ge 1$ and $\mathbf{E}Y < \infty$, show that $\lim_{n\to\infty} \mathbf{E}X_n = \mathbf{E}X$.

Theorem 9.5 (**Lévy Continuity Theorem, Special Case**). *Let $X, X_1, X_2, \ldots$ be real-valued random variables (possibly on different sample spaces). The following are equivalent.*

- *For every $t \in \mathbb{R}$, $\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$.*
- *$X_1, X_2, \ldots$ converges in distribution to X.*

Proof. The second condition implies the first by Exercise 9.4(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 8.6, 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 2.40 (and Proposition 8.4(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 8.2(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 9.2. □

Remark 9.6. 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 9.7 (**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 $\limsup_{n\to\infty}(1 - \mu_{X_n}([-m, m])) \leq \varepsilon$.)
(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 9.3 to get from (ii) to other conditions.)

10. Appendix: Moment Generating Functions

Exercise 10.1. 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 $s = \log(x) - n$.)

For any $w \in \mathbb{R}^k$ define

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

Define now

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

Lemma 10.2. *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 $\beta \in \mathbb{R}$

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

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

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

So, if

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

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

then $|X_\varepsilon| \le Y$ for any $0 < \varepsilon < \delta < 1$. We then conclude by the Dominated Convergence Theorem 2.40 that

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

$$= \int_{\mathbb{R}^n} t_1(x) h(x) \exp\Big(\sum_{i=1}^k w_i t_i(x) \Big) d\mu(x).$$

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. □

Theorem 10.3 (**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 10.2 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 10.2 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,

$$\left| e^{itx} \Big(e^{ihx} - \sum_{k=0}^n \frac{(ihx)^n}{n!} \Big) \right| = \left| e^{ihx} - \sum_{k=0}^n \frac{(ihx)^k}{k!} \right| \le \frac{|hx|^{n+1}}{(n+1)!}.$$

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 10.2, 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 9.6. $\square$

11. 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 $\theta \in \Theta$

$\mathbf{P}_\theta$ 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