MATH 541B, GRADUATE MATHEMATICAL STATISTICS II, FALL 2023

STEVEN HEILMAN

Contents

Date: December 2, 2023

1. Review of Probability Theory

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

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

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

Definition 1.3 (**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 $A_i \cap A_j = \emptyset$.

The following properties follow from the above definitions.

Proposition 1.4. *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 1.5 (**De Morgan's Laws**)**.** Let $A_1, A_2, \ldots$ be sets in some universe Ω. Then

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

Definition 1.6. 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 1.7. 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 1.8 (**Multiplication Rule**). *Let n be a positive integer. Let $A_1, \ldots, A_n$ be sets in some sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Assume that $\mathbf{P}(A_i) > 0$ for all $i \in \{1, \ldots, n\}$. Then*

$$\mathbf{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_2 \cap A_1) \cdots \mathbf{P}(A_n| \cap_{i=1}^{n-1} A_i).$$

Theorem 1.9 (**Total Probability Theorem**). *Let $A_1, \ldots,$ be disjoint events in a sample space Ω. That is, $A_i \cap A_j = \emptyset$ whenever $i, j \geq 1$ satisfy $i \neq j$. Assume also that $\cup_{i=1}^{\infty} A_i = \Omega$. Let $\mathbf{P}$ be a probability law on Ω. Then, for any event $B \subseteq \Omega$, we have*

$$\mathbf{P}(B) = \sum_{i=1}^{\infty} \mathbf{P}(B \cap A_i) = \sum_{i=1}^{\infty} \mathbf{P}(A_i)\mathbf{P}(B|A_i).$$

Proposition 1.10 (**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

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

Definition 1.11 (**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 1.12. 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 1.13 (**Probability Mass Function**)**.** Let X be a discrete random variable on a sample space Ω, so that $X : \Omega \to \mathbb{R}$. 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 1.14 (**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 $B \subseteq \mathbb{R}^n$.

We now give descriptions of some commonly encountered random variables.

Definition 1.15 (**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 1.16 (**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 1.17 (**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 1.18 (**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 1.19 (**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 1.20 (**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 1.21. 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 1.22. 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 1.23 (**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 1.24 (**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 1.25. *Let $\lambda > 0$. For each positive integer n, let $\lambda_n > 0$. Assume that $\lim_{n\to\infty} \lambda_n = \lambda$. Then*

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

Remark 1.26. 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 1.27. 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 1.28 (**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.

Definition 1.29 (**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

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

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

Definition 1.30 (**Cauchy Distribution**)**.** Define the (centered) **Cauchy distribution** to be the random variable with the probability density function

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

Definition 1.31 (**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 1.32 (**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 1.33. 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 1.34. 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

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

Proposition 1.35. *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 1.36 (**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 1.37. *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 1.38 (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.

Definition 1.39 (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 1.40. 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 $-\infty \leq a < b \leq \infty$ and $-\infty \leq c < d \leq \infty$, we have

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

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

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

Comparing this formula with Definition 1.11, 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.

Definition 1.41. 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,

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

Exercise 1.42. Let X, Y be random variables with joint PDF $f_{X,Y}$. Let $a, b \in \mathbb{R}$. Using Definition 1.41, show that $\mathbf{E}(aX + bY) = a\mathbf{E}X + b\mathbf{E}Y$.

Definition 1.43 (Joint Density Function). We say that random variables $X_1, \ldots, X_n$ have **joint density function** $f\colon \mathbb{R}^n \to [0,\infty)$ if $\int_{\mathbb{R}^n} f(x)dx = 1$, and if

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

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

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

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

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

And so on.

Exercise 1.44. 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

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

We defined independence of random variables in Definition 1.14. Below is an equivalent definition (the equivalence is beyond the scope of this course).

Definition 1.45 (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 1.46. 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 1.47. 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 1.48. 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

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

Proposition 1.49. *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*

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

Proposition 1.50. *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 1.51. *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 1.52 (Covariance). Let X and Y be random variables with finite variances. We define the **covariance** of X and Y, denoted $\mathrm{cov}(X, Y)$, by

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

Remark 1.53. By the Cauchy-Schwarz inequality (see Theorem 1.71), we have

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

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

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

The covariance of X and Y is meant to measure whether or not X and Y are related somehow. The covariance of two random variables can be any real number. In order to more accurately measure how two random variables are "related" to each other, it is natural to divide the covariance by the product of the standard deviations, i.e. the right side of Remark 1.53.

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

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

From Remark 1.53, 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 1.71). 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 1.55. Let $X_1,\ldots,X_n$ be random variables. Then

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

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

Definition 1.56. 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 1.57. *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.*

Remark 1.58. Proposition 1.57 implies that facts from Proposition 1.10 apply also to conditional probabilities. For example, using the notation of Proposition 1.57, we have $\mathbf{P}(A \cup C|B) \leq \mathbf{P}(A|B) + \mathbf{P}(C|B)$.

Definition 1.59 (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

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

Definition 1.60 (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 1.40, note that $\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = 1$. So, $f_{X|Y}(x|y)$ is a probability distribution function.

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

Theorem 1.61 (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*

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

Exercise 1.62. 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 1.63 (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 1.62 with $y := \mathbf{E}X$.) Deduce the **triangle inequality**:

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

Exercise 1.64 (**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 1.65. *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 1.64,

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

□

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

Definition 1.66 (**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 1.67. By Jensen's Inequality, if $\mathbf{E}X^2 < \infty$, then $\mathbf{E}\,|X| < \infty$, so $\mathbf{E}X \in \mathbb{R}$.

Exercise 1.68. 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 1.65 gives:

Corollary 1.69 (**Chebyshev's Inequality**)**.** *Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable with $\mathbf{E}X^2 < \infty$. Then*

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

(By Exercise 1.63, $\mathbf{E}X \in \mathbb{R}$.)

Corollary 1.65 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 1.70 (**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 1.71 (**Hölder's Inequality**)**.** *Let $X, Y: \Omega \to \mathbb{R}$ 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 1.72 (**Triangle Inequality**)**.** *Let $X, Y: \Omega \to \mathbb{R}$ 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. □

Let X, Y be independent random variables. From Proposition 1.57, 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 1.73. 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 1.74 (**Convolution on the integers**)**.** Let $g, h: \mathbb{Z} \to \mathbb{R}$ be functions. The **convolution** of g and h, denoted $g * h$, is a function $g * h: \mathbb{Z} \to \mathbb{R}$ defined by

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

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 1.75 (**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 1.76. *Let X, Y be two continuous independent random variables. Assume that f_Y is a continuous function. Then*

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

Exercise 1.77 (**Convolution is Associative**)**.** Let $g, h, d\colon \mathbb{R} \to \mathbb{R}$. Then for any $t \in \mathbb{R}$,

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

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 1.6. 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.

1.1. **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.

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

Definition 1.78 (**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 $Y\colon \Omega \to \mathbb{R}$ if

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

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

Definition 1.79 (**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 $Y\colon \Omega \to \mathbb{R}$ if: for all $\varepsilon > 0$,

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

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

Definition 1.80 (**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 $s \mapsto \mathbf{P}(Y \leq s)$ is continuous at $s = t$,

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

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

Definition 1.81 (**Convergence in L_p**)**.** Let $0 < p \leq \infty$. We say that random variables $Y_1, Y_2, \ldots \colon \Omega \to \mathbb{R}$ **converge in L_p** to $Y \colon \Omega \to \mathbb{R}$ 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 $\|X\|_\infty := \inf\{c > 0 \colon \mathbf{P}(|X| \leq c) = 1\}$.)

Exercise 1.82. Let $Y_1, Y_2, \ldots \colon \Omega \to \mathbb{R}$ be random variables that converge almost surely to a random variable $Y \colon \Omega \to \mathbb{R}$. 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
$$A_{n,\varepsilon} := \bigcup_{m=n}^{\infty} \{\omega \in \Omega \colon \ |Y_m(\omega) - Y(\omega)| > \varepsilon\}.$$
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 1.83. Let $0 < p \leq \infty$. Show that, if $Y_1, Y_2, \ldots \colon \Omega \to \mathbb{R}$ converge to $Y \colon \Omega \to \mathbb{R}$ in L_p, then $Y_1, Y_2, \ldots$ converges to Y in probability.

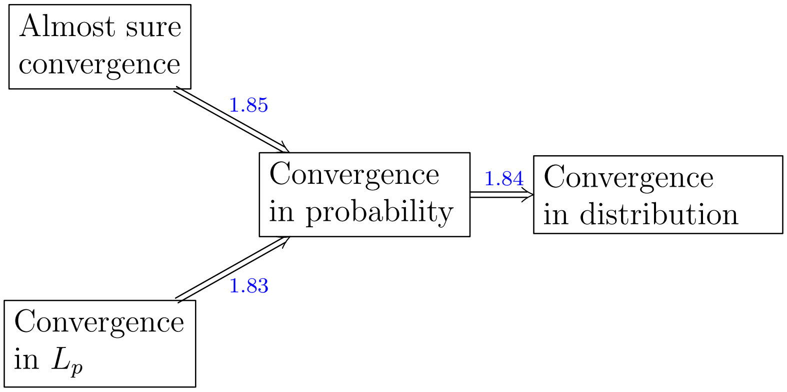
Then, show that the converse is false.

Exercise 1.84. Suppose random variables $Y_1, Y_2, \ldots \colon \Omega \to \mathbb{R}$ converge in probability to a random variable $Y \colon \Omega \to \mathbb{R}$. Prove that $Y_1, Y_2, \ldots$ converge in distribution to Y.

Then, show that the converse is false.

Exercise 1.85. 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 1.86. The following table summarizes our different notions of convergence of random variables, i.e. the following table summarizes the implications of Exercises 1.83, 1.84 and 1.85.



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.

Theorem 1.87 (**Weak Law of Large Numbers**)**.** *Let $X_1, \ldots, X_n$ be independent identically distributed random variables. Assume that $\mu := \mathbf{E}X_1$ is finite. Then for any $\varepsilon > 0$*

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

Theorem 1.88 (**Strong Law of Large Numbers**)**.** *Let $X_1, \ldots, X_n$ be independent identically distributed random variables. Assume that $\mu := \mathbf{E}X_1$ is finite. Then*

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

Remark 1.89. A Monte Carlo simulation takes n independent samples from some random distribution and then sums the sample results and divides by n. The Strong Law of Large Numbers guarantees that this averaging procedure converges to the average value as n becomes large.

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

More formally, let $X_1, X_2, \ldots \colon \Omega \to \mathbb{R}$ 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 1.48. 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 1.69, 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 1.90 (**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 $-\infty \leq a \leq \infty$,

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

Remark 1.91. 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 1.92. 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 1.93 (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 1.94 (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 $\sigma := \sqrt{\mathrm{var}(X_1)}$. 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?

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 1.95. Let $X_1, X_2, \ldots\colon \Omega \to \mathbb{R}$ be i.i.d. with $\mathbf{E}|X_1| = \infty$. Then $\mathbf{P}(|X_n| > n$ 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 1.96 (Second Borel-Cantelli Lemma). Let $A_1, A_2, \ldots$ be independent events with $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$. Then $\mathbf{P}(A_n$ 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 1.97 (**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 1.98 (**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 1.99. *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 1.100. 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 1.101 (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 1.102. 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 1.103. 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 1.104. Let $X_1, X_2, \ldots \colon \Omega \to \mathbb{R}$ 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 1.105.

- *(Slutsky's Theorem) Let $X_1, X_2, \ldots \colon \Omega \to \mathbb{R}$ be random variables that converge in distribution to $X \colon \Omega \to \mathbb{R}$. Let $c \in \mathbb{R}$. Let $Y_1, Y_2, \ldots \colon \Omega \to \mathbb{R}$ 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, $X_1Y_1, X_2Y_2, \ldots$ converges in distribution to cX.*
- *Let $X_1, X_2, \ldots \colon \Omega \to \mathbb{R}$ be random variables that converge in distribution 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 distribution to $f(X)$.*

Exercise 1.106. 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 1.107. 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.)

2. Review of Statistics

2.1. **Exponential Families.**

A basic problem in statistics is to fit data to an unknown probability distribution. As in Exercise 1.106, we might have a list of numbers, and we known these numbers follow some Gaussian distribution, but we might not know the mean and variance of this Gaussian. We then want to infer the mean and variance from the data. In this example, there are two unknown parameters. In order to generalize this problem, we introduce exponential families. Exponential families provide a general class of distributions with a given number of unknown parameters.

Definition 2.1 (**Exponential Families).** Let n, k be a positive integers and let μ be a measure on $\mathbb{R}^n$. Let $t_1, \ldots, t_k \colon \mathbb{R}^n \to \mathbb{R}$. Let $h \colon \mathbb{R}^n \to [0, \infty]$ so that h is not identically zero. For any $w = (w_1, \ldots, w_k) \in \mathbb{R}^k$, define

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

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

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

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

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

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

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

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

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

Remark 2.2. If $w \colon \Theta \to \mathbb{R}^k$ has an inverse function, then the corresponding k-parameter exponential family can be written in canonical form.

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

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

and we can then interpret

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

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

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

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

and we can then interpret

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

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

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

Example 2.3. Let us see how to phrase Exercise 1.106 using a two parameter exponential family. We write a Gaussian density of mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$ as

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

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

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

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

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

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

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

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

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

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

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

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

Lemma 2.4. *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.*

Lemma 2.4 is proven in Lemma 12.2 below.

Theorem 2.5 (Dominated Convergence Theorem). *Let $X_1, X_2, \ldots \colon \Omega \to \mathbb{R}$ 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*

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

Theorem 2.6. ***(Dominated Convergence Theorem)*** *Let E be a measurable set. Let $\{f_n\}$ be a sequence of measurable functions such that $|f_n| \leq g$, g integrable. If $f = \lim f_n$ exists then $\lim \int_E f_n$ exists, f is integrable on E, and*

$$\int_E f d\mu = \lim \int_E f_n d\mu$$

Corollary 2.7. *Let $\varepsilon > 0$. Let $X \colon \Omega \to \mathbb{R}$ be a random variable such that $\mathbf{E}e^{wX} < \infty$ for all $w \in (-\varepsilon, \varepsilon)$. Then, for any integer $n \geq 1$, $\mathbf{E}X^n$ exists and*

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

Exercise 2.8. Let $X \colon \Omega \to \mathbb{R}^n$ be a random variable with the **standard Gaussian distribution**:

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

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.)

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

2.2. **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 2.9 (**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 1.106, 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 2.10 (**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 2.11. 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 2.12. 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 2.13. 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 $\sigma := \sqrt{\mathrm{var}(X_1)} < \infty$. Let $\overline{X}$ be the sample mean and let S be the sample standard deviation of the random sample. Show the following

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

If we divided by n instead of $n - 1$ in the definition of S, then the second part of the above exercise would not hold. Since $\mathbf{E}S^2$ agrees with the variance of X, we say that S^2 is unbiased. We will discuss this concept more in Section 2.4.

Exercise 2.14. 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 $t = \mathbf{E}X$.

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 2.15. *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.*

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

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

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 2.16. *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 2.17. If $X_1, \ldots, X_{n+1}$ 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 2.15).

Exercise 2.18. 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 2.19 (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 $X_{(1)} := \min_{1\leq i\leq n} X_i$, and for any $2 \leq i \leq n$, inductively define

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

so that

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

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 $i \geq 1$, define $p_i := \mathbf{P}(X \leq x_i)$. Show that, for any $1 \leq i, j \leq n$,

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

  (Hint: Let Y be the number of indices $1 \leq j \leq n$ such that $X_j \leq x_i$. 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 $1 \leq j \leq n$, $F_{X_{(j)}}$ 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 $\mathbf{P}(X_{(j)} \leq x) = \sum_{k=j}^{n} \binom{n}{k} F_X(x)^k (1-F_X(x))^{n-k}$, where $F_X(x) := \mathbf{P}(X \leq x)$ for any $x \in \mathbb{R}$.)

- Let X be a random variable uniformly distributed in $[0,1]$. For any $1 \leq j \leq n$, 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 $1 \leq j \leq n$ such that $X_j \leq a$. Let V be the number of indices $1 \leq j \leq n$ such that $a < X_j \leq b$. Show that the vector $(U, V, n-U-V)$ is a multinomial random variable, so that for any nonnegative integers u, v with $u + v \leq n$, 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 $1 \leq i, j \leq n$,

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

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).

From Examples 2.11 and 2.12 and Exercise 2.13, 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 2.20 (**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 $n \to \infty$.

Theorem 2.21 (**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 2.22 (**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 $n \to \infty$.

Let $m > 2$ be an integer. Theorem 2.22 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 $\sigma^m \frac{1}{m!} f^{(m)}(\theta)$.

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.

2.3. Data Reduction.

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

Definition 2.23 (**Sufficient Statistic**). Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a distribution f where $f \in \{f_\theta \colon \theta \in \Theta\}$ is a family of PDFs or PMFs (such as an exponential family). Let $t\colon \mathbb{R}^n \to \mathbb{R}^k$, so that $Y := t(X_1, \ldots, X_n)$ is a statistic. We say that Y is a **sufficient statistic** for θ if, for every $y \in \mathbb{R}^k$ and for every $\theta \in \Theta$, the conditional distribution of $(X_1, \ldots, X_n)$ given $Y = y$ (with respect to probabilities given by f_θ) does not depend on θ. That is, Y provides sufficient information to determine θ from $X_1, \ldots, X_n$.

Definition 2.24 (**Minimal Sufficient Statistic**). Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a family $\{f_\theta \colon \theta \in \Theta\}$ of joint probability density functions, or joint probability mass functions. Let $t\colon \mathbb{R}^n \to \mathbb{R}^k$, so that $Y := t(X_1, \ldots, X_n)$ is a statistic. Assume that Y is sufficient for θ. Then Y is **minimal sufficient** for θ if, for every statistic $Z\colon \Omega \to \mathbb{R}^m$ that is sufficient for θ, there exists a function $r\colon \mathbb{R}^m \to \mathbb{R}^k$ such that $Y = r(Z)$.

Theorem 2.25. *Suppose $(X_1, \ldots, X_n)$ is a random sample of size n from a family $\{f_\theta \colon \theta \in \Theta\}$ of joint probability density functions or joint probability mass functions. (In the case of probability mass functions, we also assume that the set $\cup_{\theta\in\Theta}\{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$ is countable.) Let $t\colon \mathbb{R}^n \to \mathbb{R}^m$ and define $Y := t(X_1, \ldots, X_n)$. When $\{f_\theta \colon \theta \in \Theta\}$ are joint probability density functions, suppose the following condition holds for every $x, y \in \mathbb{R}^n$, and when $\{f_\theta \colon \theta \in \Theta\}$ are joint probability mass functions, suppose the following condition holds for every x, y in the set $\cup_{\theta\in\Theta}\{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$.*

$$\begin{gathered}\exists\, c(x,y) \in \mathbb{R} \text{ that does not depend on } \theta \text{ such that} \\ f_\theta(x) = c(x,y) f_\theta(y) \quad \forall\, \theta \in \Theta \\ \text{if and only if } t(x) = t(y).\end{gathered}$$

Then Y is minimal sufficient.

Exercise 2.26. 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

$$t = s(u).$$

Remark 2.27. If a minimal sufficient statistic exists, it is unique up to an invertible transformation. To see this, let $Y\colon \Omega \to \mathbb{R}^n$ and let $Z\colon \Omega \to \mathbb{R}^m$ be minimal sufficient statistics. By minimality of Y, there exists $r\colon \mathbb{R}^m \to \mathbb{R}^n$ such that $Y = r(Z)$. By minimality of Z, there exists $s\colon \mathbb{R}^n \to \mathbb{R}^m$ such that $Z = s(Y)$. Composing each of these identities with each other, we getj

$$Y = r(s(Y)), \qquad Z = s(r(Z)).$$

That is, $r \circ s$ is the identity map on the range of Y, and $s \circ r$ is the identity map on the range of Z. That is, Y and Z are each the invertible image of each other.

The uniqueness of the minimal sufficient statistic is nice, since it implies that (up to an invertible map), there is at most one way to reduce the data at hand when we are trying to determine the parameter θ that fits our data.

Proposition 2.28 (**Existence of Minimal Sufficient Statistic**). *Suppose $X_1, \ldots, X_n$ is a random sample of size n from a distribution f where $f \in \{f_\theta\colon \theta \in \Theta\}$ is a family of probability density functions, or a family of probability mass functions. (In the case of probability mass functions, we also assume that the set $\cup_{\theta\in\Theta}\{x \in \mathbb{R}^n\colon f_\theta(x) > 0\}$ is countable.) Then there exists a statistic Y that is minimal sufficient for θ.*

Minimal sufficient statistics provide sufficient information to estimate a parameter θ in a family of distributions $\{f_\theta\colon \theta \in \Theta\}$. However, even a minimal sufficient statistic can have excess "information." For example, we saw in Proposition 2.28 that a minimal sufficient statistic can have infinitely many nontrivial components in its range. It would be desirable to come up with statistics that contain as little unnecessary information as possible, while still being minimal sufficient. In order to accomplish this task, we first define what we mean by "excess information" of a statistic.

Definition 2.29 (**Ancillary Statistic**). Suppose $X_1, \ldots, X_n$ is a random sample of size n from a distribution f where $f \in \{f_\theta\colon \theta \in \Theta\}$ is a family of distributions. A statistic $Y = t(X_1, \ldots, X_n)$, $t\colon \mathbb{R}^n \to \mathbb{R}^m$ is **ancillary** for θ if the distribution of Y does not depend on θ.

Example 2.30. Let $X_1, \ldots, X_n$ be a random sample of size n from the location family for the Cauchy distribution:

$$f_\theta(x) := \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1 + (x_i - \theta)^2}, \qquad \forall\, x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad \forall\, \theta \in \mathbb{R}.$$

Then the order statistics $X_{(1)} \leq \cdots \leq X_{(n)}$ are minimal sufficient for θ. Sufficiency follows by the Factorization Theorem 2.48 since, if $t(x) := (x_{(1)}, \ldots, x_{(n)})$, then $f_\theta(t(x)) = f_\theta(x)$. Minimal sufficiency follows from Theorem 2.25, since if $x, y \in \mathbb{R}^n$ are fixed, then the following ratio is constant in θ

$$\frac{f_\theta(x)}{f_\theta(y)} = \frac{\prod_{i=1}^{n} \frac{1}{1+(x_i-\theta)^2}}{\prod_{i=1}^{n} \frac{1}{1+(y_i-\theta)^2}} = \frac{\prod_{i=1}^{n}[1 + (y_i - \theta)^2]}{\prod_{i=1}^{n}[1 + (x_i - \theta)^2]},$$

only when $t(x) = t(y)$. To see this, note that the top and bottom are each polynomials in θ, and these polynomials must be a constant multiple of each other, so their (complex) roots must be identical (counting multiplicities), and these roots are $\theta = x_i \pm \sqrt{-1}$, $\theta = y_i \pm \sqrt{-1}$ respectively ($i = 1, \ldots, n$), so that $t(x) = t(y)$.

Even though the order statistics $(X_{(1)}, \ldots, X_{(n)})$ are minimal sufficient for θ in this case, they certainly seem to contain a lot of extraneous information about θ. Indeed, the statistic $X_{(n)} - X_{(1)}$ is ancillary. To see this, let $Z_1, \ldots, Z_n$ be independent Cauchy random variables, i.e. they each have density $\frac{1}{\pi}\frac{1}{1+x^2}$ for all $x \in \mathbb{R}$. Then $X_i = Z_i + \theta$ for all $1 \leq i \leq n$, so that $X_{(i)} = Z_{(i)} + \theta$ for all $1 \leq i \leq n$, so that $X_{(n)} - X_{(1)} = Z_{(n)} - Z_{(1)}$, and the last expression does not depend on θ.

Definition 2.31 (**Complete Statistic**). Suppose $X_1, \ldots, X_n$ is a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. Let $t \colon \mathbb{R}^n \to \mathbb{R}^m$. A statistic $Y = t(X_1, \ldots, X_n)$ is **complete** for $\{f_\theta \colon \theta \in \Theta\}$ if the following holds:

$$\text{For any } f \colon \mathbb{R}^m \to \mathbb{R} \text{ such that } \mathbf{E}_\theta f(Y) = 0 \quad \forall\, \theta \in \Theta, \qquad \text{it holds that } f(Y) = 0.$$

(When we assume that $\mathbf{E}_\theta f(Y)$ can be defined, we also assume that $\mathbf{E}_\theta |f(Y)| < \infty$ for all $\theta \in \Theta$.)

Remark 2.32. From our discussion above, we see that a nonconstant complete statistic is not ancillary. (If Y is ancillary, then there is a constant $c \in \mathbb{R}$ such that $\mathbf{E}_\theta(Y - c) = 0$ for all $\theta \in \Theta$, and if Y is also complete, we then have $Y - c = 0$, so that $Y = c$.) Also, a complete statistic may not be sufficient. Consider for example a statistic that is constant.

Remark 2.33. Unfortunately, a complete sufficient statistic might not exist.

Exercise 2.34. Give an example of a statistic Y that is complete and nonconstant, but such that Y is not sufficient.

Theorem 2.35 (**Bahadur's Theorem**). *If Y is a complete sufficient statistic for a family $\{f_\theta \colon \theta \in \Theta\}$ of joint probability densities or joint probability mass functions, then Y is a minimal sufficient statistic for θ. (In the case of probability mass functions, we also assume that the set $\cup_{\theta\in\Theta}\{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$ is countable.)*

Remark 2.36. So, by Remark 2.27, a complete sufficient statistic is unique, up to an invertible map. Also, by Example 2.30, the converse of Bahadur's Theorem is false.

The following theorem says that complete sufficient statistics have no ancillary information, unlike the minimal sufficient statistics, as we saw in Example 2.30.

Theorem 2.37 (**Basu's Theorem**). *If Y is a complete sufficient statistic for $\{f_\theta \colon \theta \in \Theta\}$, and if Z is ancillary for θ, then for all $\theta \in \Theta$, Y and Z are independent with respect to f_θ.*

Sufficient and ancillary statistics were introduced by Fisher in 1920. Complete and minimal sufficient statistics were studied in the mid 1900s by Bahadur, Halmos, and Savage, and Lehmann and Scheffé.

Above, we have typically focused on families of probability density functions or probability mass functions, in order to avoid use of measure theory. However, many of the above theorems naturally generalize to the setting of a dominated family of functions.

Definition 2.38 (**Dominated Family**)**.** Let $\Theta \subseteq \mathbb{R}^m$. Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of functions so that $f_\theta \colon \mathbb{R}^n \to [0, \infty)$ for all $\theta \in \Theta$. We say that $\{f_\theta \colon \theta \in \Theta\}$ is a **dominated family** if there exists a measure μ on $\mathbb{R}^m$ such that $\mathbf{P}_\theta$ is absolutely continuous with respect to μ, for all $\theta \in \Theta$.

For example, a family of probability density functions is absolutely continuous with respect to Lebesgue measure. And a family of probability mass functions is absolutely continuous with respect to a counting measure, if $\cup_{\theta \in \Theta}\{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$ is countable.

We can then restate the Factorization Theorem and its Corollaries for dominated families.

Theorem 2.39 (**Factorization Theorem**)**.** *Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a dominated family $\{f_\theta \colon \theta \in \Theta\}$ that is dominated by a measure μ on $\mathbb{R}^n$. That is, $f_\theta \colon \mathbb{R}^n \to [0, \infty)$ for all $\theta \in \Theta$. Let $t \colon \mathbb{R}^n \to \mathbb{R}^k$, so that $Y := t(X_1, \ldots, X_n)$ is a statistic. Then Y is sufficient for θ if and only if there exist nonnegative functions $\{g_\theta \colon \theta \in \Theta\}$, $h \colon \mathbb{R}^n \to [0, \infty)$, $g_\theta \colon \mathbb{R}^k \to [0, \infty)$, such that*

$$f_\theta(x) = g_\theta(t(x))h(x), \qquad \forall\, \theta \in \Theta, \quad \textit{for a.e. } x \textit{ with respect to } \mu.$$

Theorem 2.40. *Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a dominated family $\{f_\theta \colon \theta \in \Theta\}$ that is dominated by a measure μ on $\mathbb{R}^n$. Let $t \colon \mathbb{R}^n \to \mathbb{R}^m$ and define $Y := t(X_1, \ldots, X_n)$. Suppose the following condition holds for a.e. $x, y \in \mathbb{R}^n$ with respect to μ:*

$$\exists\, c(x,y) \in \mathbb{R} \text{ that does not depend on } \theta \text{ such that}$$

$$f_\theta(x) = c(x,y) f_\theta(y) \quad \forall\, \theta \in \Theta$$

$$\text{if and only if } t(x) = t(y).$$

Then Y is minimal sufficient.

Proposition 2.41 (**Existence of Minimal Sufficient Statistic**)**.** *Let $X = (X_1, \ldots, X_n)$ be a random sample of size n from a dominated family $\{f_\theta \colon \theta \in \Theta\}$ that is dominated by a measure μ on $\mathbb{R}^n$. Suppose the set $\{f_\theta \colon \theta \in \Theta\}$ has a countable dense set with respect to the total variation metric $d(f_\theta, f_{\theta'}) = \sup_{B \subseteq \mathbb{R}^n} |\mathbf{P}_\theta(B) - \mathbf{P}_{\theta'}(B)|$. Then there exists a statistic Y that is minimal sufficient for θ.*

To see the original proof, read Theorem 6.1 in "Completeness, Similar Regions, and Unbiased Estimation-Part I" by Lehmann and Scheffé.

2.4. **Estimation of Parameters.**

A basic problem in statistics is to fit data to an unknown probability distribution. As in Exercise 1.106, 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 $X_1, \ldots, X_n$ 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 2.42. In Exercise 1.106 we have a random sample $X_1, \ldots, X_{20}$ from a Gaussian distribution with unknown mean and variance. We denote the unknown Gaussians as

$$\{f_\theta \colon \theta \in \Theta\} = \{f_{\mu,\sigma}(x) \colon (\mu,\sigma) \in \mathbb{R}^2,\, \mu \in \mathbb{R},\, \sigma > 0\} = \Big\{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \colon \mu \in \mathbb{R},\, \sigma > 0\Big\}.$$

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 2.43. 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 $t \colon \mathbb{R}^n \to \mathbb{R}^k$ and let $Y := t(X_1, \ldots, X_n)$ 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 2.13 that the sample mean and sample variance are unbiased estimates of the mean and variance, respectively.

Definition 2.44 (Consistency). Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of distributions. Let $Y_1, Y_2, \ldots$ be a sequence of estimators of $g(\theta)$ where $g \colon \Theta \to \mathbb{R}^k$. We say that $Y_1, Y_2, \ldots$ is **consistent** for $g(\theta)$ if, for any $\theta \in \Theta$, $Y_1, Y_2, \ldots$ converges in probability to the constant value $g(\theta)$, with respect to the probability distribution f_θ.

Typically, we will take Y_n to be a function of a random sample of size n, for all $n \geq 1$.

Definition 2.45 (Method of Moments). Let $g \colon \Theta \to \mathbb{R}^k$. Suppose we want to estimate $g(\theta)$ for any $\theta \in \Theta$. Suppose there exists $h \colon \mathbb{R}^j \to \mathbb{R}^k$ such that

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

Then the estimator

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

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

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 2.51 below, to get a good estimate of the parameter θ.

Definition 2.46 (**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 $X_1, \ldots, X_n$.

Example 2.47. 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 2.48 (**Factorization Theorem**)**.** *Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a family $\{f_\theta : \theta \in \Theta\}$ 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 $\cup_{\theta\in\Theta}\{x \in \mathbb{R}^n : f_\theta(x) > 0\}$ is countable.) Let $t: \mathbb{R}^n \to \mathbb{R}^k$, 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 $\{g_\theta : \theta \in \Theta\}$, $h: \mathbb{R}^n \to [0, \infty)$, $g_\theta : \mathbb{R}^k \to [0, \infty)$, such that*

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

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

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 $B \subseteq \cup_{i=1}^\infty B_i$.

Exercise 2.49 (**Conditional Expectation as a Random Variable**)**.** Let $X, Y, Z : \Omega \to \mathbb{R}$ be discrete or continuous random variables. Let A be the range of Y. Define $g: A \to \mathbb{R}$ 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 $\{(x, y) \in \mathbb{R}^2 : x \geq 0, y \geq 0, x + y \leq 1\}$. 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 $f : \mathbb{R} \to \mathbb{R}$ is uniquely minimized when $t = \mathbf{E}X$. A similar minimizing property holds for conditional expectation. Let $h : \mathbb{R} \to \mathbb{R}$. Show that the quantity $\mathbf{E}(X - h(Y))^2$ is minimized among all functions $h : \mathbb{R} \to \mathbb{R}$ 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 2.50 (UMVU). Let $X_1, \ldots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta : \theta \in \Theta\}$. Let $g : \Theta \to \mathbb{R}$. Let $t : \mathbb{R}^n \to \mathbb{R}$ 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

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

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

Theorem 2.51 (Rao-Blackwell). *Let Z be a sufficient statistic for $\{f_\theta\colon \theta\in\Theta\}$ and let Y be an estimator for $g(\theta)$. Define $W:=\mathbf{E}_\theta(Y|Z)$. (Since Z is sufficient for θ, W does not depend on θ by Exercise 2.54, i.e. W is a well-defined function of the random sample but not an explicit function of θ.) Let $\theta\in\Theta$ with $r(\theta,Y)<\infty$ and such that $\ell(\theta,y)$ is convex in $y\in\mathbb{R}$. Then*

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

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

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

Definition 2.52. 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 2.53 (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 2.26) 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 2.54. 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 2.55. By Exercise 2.49, 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 2.51.

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 2.56 (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 10.9 applies.) We also have another equivalent definition:

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

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

Example 2.57. 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

$$I(\theta)=\mathrm{Var}_\theta\Big(\frac{d}{d\theta}\frac{-(X-\theta)^2}{2\sigma^2}\Big)=\frac{1}{\sigma^4}\mathrm{Var}_\theta(X-\theta)=\frac{1}{\sigma^2}.$$

Proposition 2.58. *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*

$$I_{(X,Y)}(\theta)=I_X(\theta)+I_Y(\theta),\qquad \forall\,\theta\in\Theta.$$

Our primary interest in information is the following inequality. Theorem 2.59 gives a lower bound on the variance of unbiased estimators of θ.

Theorem 2.59 (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 10.9 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 2.60. 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 2.58 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 2.59 says

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

For a one-parameter family of distributions, the equality case of Theorem 2.59 allows us to find a UMVU for θ. To find such an estimator, we look for affine functions of $\frac{d}{d\theta}\log f_\theta(X)$.

Let $X_1,\ldots,X_n$ be a random sample of size n from a family of distributions $\{f_\theta\colon \theta\in\Theta\}$. So, we denote the joint distribution of $X_1,\ldots,X_n$ as

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

If we have data $x\in\mathbb{R}^n$, recall that we defined the function $\ell\colon\Theta\to[0,\infty)$

$$\ell(\theta):=\prod_{i=1}^n f_\theta(x_i)$$

and called it the **likelihood function**.

Definition 2.61 (**Maximum Likelihood Estimator**)**.** The **maximum likelihood estimator** (MLE) Y is the estimator maximizing the likelihood function. That is, $Y:=t(X)$, $t\colon\mathbb{R}^n\to\Theta$ and $t(x_1,\ldots,x_n)$ is defined to be any value of $\theta\in\Theta$ that maximizes the function

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

if this value of θ exists. A priori, the θ maximizing $\ell(\theta)$ might not exist, and it might not be unique

Remark 2.62. Maximizing the likelihood $\ell(\theta)$ is equivalent to maximizing $\log\ell(\theta)$, since log is monotone increasing.

Example 2.63. 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 1.68 that, if $z\neq\frac{1}{n}\sum_{i=1}^n x_i$, then

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

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 2.47 M_1 is sufficient for the mean, by the Rao-Blackwell Theorem 2.51 $\mathbf{E}_\theta(M_1|M_1)$ is UMVU for the mean, and $\mathbf{E}_\theta(M_1|M_1) = M_1$ by Exercise 2.49(iv)). Let

$$Y = Y_n = Y_n(X_1, \ldots, X_n) := \frac{1}{n}\sum_{j=1}^{n}\Big(X_j - \frac{1}{n}\sum_{i=1}^{n} X_i\Big)^2.$$

We also know from Proposition 2.15 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 2.57 and using Proposition 2.58

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

By the Cramér-Rao Inequality, Theorem 2.59, with $g(\sigma) = \mathbf{E}_\sigma(Y) = \sigma^2(n-1)/n$ (using Proposition 2.15), 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 2.15,

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

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 2.64. 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 1.90, $\sqrt{n}(\overline{X}_n - \theta^{-1})$ converges in distribution to a Gaussian random variable with mean 0 and variance θ^{-2} as $n \to \infty$. So, the Delta Method, Theorem 2.20, 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 2.21)

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

On the other hand, the information inequality, Theorem 2.59, says the smallest possible variance of an unbiased estimator of θ is

$$1/\mathrm{Var}\Big(\frac{n}{\theta} - \sum_{i=1}^{n} X_i\Big) = 1/(n\theta^{-2}) = \theta^2/n.$$

So, the MLE asymptotically achieves the optimal variance for an estimator of θ.

Proposition 2.65 (**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 $g(\theta)$.*

Lemma 2.66 (**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 2.67. 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.

Theorem 2.68 (**Consistency of MLE**). *Let $X_1, X_2, \ldots \colon \Omega \to \mathbb{R}^n$ 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. $x_1 \in \mathbb{R}$. (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 1.87, 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 2.66 (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 2.65 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 2.69 (**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*

$$I_{X_1}(\theta) = \mathbf{E}_\theta\Big(\frac{d}{d\theta} \log f_\theta(X_1)\Big)^2 = -\mathbf{E}_\theta \frac{d^2}{d\theta^2} \log f_\theta(X_1) > 0.$$

(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 2.70. Combining this Theorem with Proposition 2.65, under the above assumptions (and also if the variance of the MLE converges, i.e. we can apply something like Theorem 2.21), the MLE for θ achieves the asymptotically optimal variance in the Cramér-Rao Inequality, Theorem 2.59. 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 1.90 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 2.71. □

Exercise 2.71. 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.

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

Theorem 2.72 (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 1.103, 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 10.9 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 2.73 (Limiting Distribution of MLE). *Let $\{f_\theta\colon \theta\in\Theta\}$ be a family of probability density functions, so that $f_\theta\colon \mathbb{R}^n \to [0,\infty)$ $\forall\,\theta\in\Theta$. Let $X_1, X_2, \ldots$ be i.i.d. such that X_1 has density f_θ. Let $\Theta \subseteq \mathbb{R}^m$. Assume the following*

- (i) *The set $A := \{x \in \mathbb{R}^n\colon f_\theta(x) > 0\}$ does not depend on θ.*
- (ii) *For every $x \in A$, $\forall\ 1 \leq i,j \leq m$, $\frac{\partial^2 f_\theta(x)}{\partial\theta_i\partial\theta_j}$ exists and is continuous in θ.*
- (iii) *The Fisher Information $I_{X_1}(\theta)$ exists and is finite, with $\mathbf{E}_\theta\nabla_\theta\log f_\theta(X_1) = 0$ and*

$$I_{X_1}(\theta) = \mathbf{E}_\theta\Big(\frac{\partial}{\partial\theta_i}\log f_\theta(X)\cdot\frac{\partial}{\partial\theta_j}\log f_\theta(X)\Big) = -\mathbf{E}_\theta D_\theta^2\log f_\theta(X_1).$$

  (D_θ^2 denotes the matrix of iterated second order derivatives in θ.) Moreover, assume that $I_{X_1}(\theta)$ is an invertible matrix.
- (iv) *For every θ in the interior of Θ, $\forall\ 1 \leq i,j \leq m$, $\exists\ \varepsilon > 0$ such that*

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

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

3. Hypothesis Testing

In Section 2.4, 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 3.1 (**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 3.2. In Exercise 1.94, 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 3.3. 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.

3.1. 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 3.4 (**Hypothesis Test**)**.** Let H_0 be a null hypothesis. A **nonrandomized 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

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

More generally, a **randomized hypothesis test** of H_0 versus H_1 is specified by a **critical function** $\phi \colon \mathbb{R}^n \to [0,1]$. For any $x \in \mathbb{R}^n$, $\phi(x)$ denotes the probability of rejecting the null hypothesis, given that we observe data x. That is,

- If $X = x$, we accept the null hypothesis to be true with probability $1 - \phi(x)$.
- If $X = x$, we reject the null hypothesis H_0 with probability $\phi(x)$.

We could identify $\{x \in \mathbb{R}^n \colon \phi(x) = 1\}$ as a rejection region, $\{x \in \mathbb{R}^n \colon \phi(x) = 0\}$ as an acceptance region, and $\{x \in \mathbb{R}^n \colon \phi(x) \in (0,1)\}$ as a region where acceptance or rejection can each occur with some nonzero probability. The **power function** $\beta \colon \Theta \to [0,1]$ is then

$$\beta(\theta) := \mathbf{E}_\theta \phi(X), \qquad \forall\, \theta \in \Theta.$$

The case $\phi\colon \mathbb{R}^n \to \{0,1\}$ then corresponds to a nonrandomized hypothesis test, where the current definition of power function agrees with our previous definition.

Remark 3.5. Since the critical function ϕ determines the hypothesis test, we will often refer to ϕ itself as a hypothesis test.

One advantage of randomized hypothesis tests over nonrandomized tests is adjustments to the test allow a wider range of power function values.

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 3.6 (**Type II Error**)**.** A **Type II Error** for a nonrandomized 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 $\theta \in \Theta_1$.

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 3.7 (**Type I Error**)**.** A **Type I Error** for a nonrandomized 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 3.8. Let us return to Exercise 1.94. 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 $\{0 \leq \theta \leq 1/2\}$, and the alternative hypothesis H_1 is $\{1/2 < \theta \leq 1\}$. 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 3.9 (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 3.10 (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 on $\mathbb{R}^n$. Fix $k \geq 0$. Define a **likelihood ratio test** $\phi\colon \mathbb{R}^n \to [0, 1]$ to be*

$$\phi(x) := \begin{cases} 1 & , \textit{if } f_{\theta_1}(x) > k f_{\theta_0}(x) \\ 0 & , \textit{if } f_{\theta_1}(x) < k f_{\theta_0}(x) \\ \text{(unspecified)} & , \textit{if } f_{\theta_1}(x) = k f_{\theta_0}(x). \end{cases} \qquad (*)$$

Define

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

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

- *(Sufficiency) Any randomized 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* $\mathbf{P}_{\theta_0}(X \in D) = \mathbf{P}_{\theta_1}(X \in D) = 0$*.*
- *(Existence) For any* $\alpha' \in [0, 1]$*, there exists a UMP class* $\mathcal{T}$ *test with* $\alpha' = \alpha$*.*

Remark 3.11. 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 3.12. When we use Lemma 3.10 in practice, we will typically fix $\alpha \in [0, 1]$, and then define $k \geq 0$ such that $\mathbf{E}_{\theta_0}\phi(X)$ is at most α (even though Lemma 3.10 instead starts with ϕ and then defines α via ϕ).

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 ϕ. Let ϕ' another test in $\mathcal{T}$, and let $\beta'(\theta)$ be the power function of this test. By definition of ϕ, we have

$$[\phi(x) - \phi'(x)][f_{\theta_1}(x) - kf_{\theta_0}(x)] \geq 0, \qquad \forall\, x \in \mathbb{R}^n.$$

Therefore,

$$0 \leq \int_{\mathbb{R}^n} [\phi(x) - \phi'(x)][f_{\theta_1}(x) - kf_{\theta_0}(x)]dx = \beta(\theta_1) - \beta'(\theta_1) - k[\beta(\theta_0) - \beta'(\theta_0)]. \qquad (***)$$

Since ϕ has significance level α and ϕ' 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 ϕ test is UMP class $\mathcal{T}$.

We now prove necessity. Let ϕ' be a UMP class $\mathcal{T}$ test. We just showed that the ϕ 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 ϕ' is a UMP class $\mathcal{T}$ test, the ϕ' 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. Necessity follows.

We now prove existence. If $\alpha' = 1$, choose $k = 0$. Below we therefore assume $\alpha' < 1$. For any $k \in \mathbb{R}$, define

$$a(k) := \mathbf{P}_{\theta_0}(f_{\theta_1}(X) > kf_{\theta_0}(X)).$$

Note that a is a monotone decreasing function of k, with $\lim_{k\to 0^-} a(k) = 1$ and $\lim_{k\to\infty} a(k) = 0$, so we can choose $c \geq 0$ such that $\lim_{k\to c^-} a(k) \geq \alpha' \geq \lim_{k\to c^+} a(k) = a(c)$. Define then

$$\phi(x) := \begin{cases} 1 & \text{, if } f_{\theta_1}(x) > cf_{\theta_0}(x) \\ 0 & \text{, if } f_{\theta_1}(x) < cf_{\theta_0}(x) \\ \frac{\alpha' - a(c)}{\mathbf{P}_{\theta_0}(f_{\theta_1}(X) = cf_{\theta_0}(X))} & \text{, if } f_{\theta_1}(x) = cf_{\theta_0}(x). \end{cases}$$

(If $\mathbf{P}_{\theta_0}(f_{\theta_1}(X) = cf_{\theta_0}(X)) = 0$, then ϕ is well-defined with $\mathbf{P}_{\theta_0}$ probability one.) Then, using the definition of ϕ,

$$\begin{aligned}\mathbf{E}_{\theta_0}\phi(X) &= \mathbf{P}_{\theta_0}(f_{\theta_1}(X) > cf_{\theta_0}(X)) + \frac{\alpha' - a(c)}{\mathbf{P}_{\theta_0}(f_{\theta_1}(X) = cf_{\theta_0}(X))}\mathbf{P}_{\theta_0}(f_{\theta_1}(X) = cf_{\theta_0}(X))\\ &= a(c) + (\alpha' - a(c)) = \alpha'.\end{aligned}$$

(If $\mathbf{P}_{\theta_0}(f_{\theta_1}(X) = kf_{\theta_0}(X)) = 0$, then $a(c) = \alpha'$ and the fraction term is zero.)

□

Example 3.13. 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 3.10 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 necessity part of Lemma 3.10.

Evidently, in order to get a UMP test with significance level other than $\{0, 1/4, 3/4, 1\}$, we need to use a randomized hypothesis test. For example, to get a UMP test with significance level 1/8, we could use $\phi\colon \mathbb{R} \to [0, 1]$ defined by

$$\phi(x) := \begin{cases} 1 & \text{, if } f_{\theta_1}(x) > (9/4)f_{\theta_0}(x) \\ 0 & \text{, if } f_{\theta_1}(x) < (9/4)f_{\theta_0}(x) \\ 1/2 & \text{, if } f_{\theta_1}(x) = (9/4)f_{\theta_0}(x) \end{cases} \qquad = \begin{cases} 0 & \text{, if } x \neq 2 \\ 1/2 & \text{, if } x = 2. \end{cases}$$

Then ϕ is UMP by Lemma 3.10 with significance level

$$\mathbf{E}_{\theta_0}\phi(X) = \mathbf{P}_{\theta_0}(X = 2)\phi(2) = \mathbf{P}_{1/2}(X = 2)(1/2) = (1/4)(1/2) = 1/8.$$

Exercise 3.14. 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 $t \in \mathbb{R}$.)

Corollary 3.15. *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. Assume that $\mathbf{P}_{\theta_0} \neq \mathbf{P}_{\theta_1}$. Let $\phi_\alpha \colon \mathbb{R}^n \to [0,1]$ be a likelihood ratio test with significance level $\alpha \in (0,1)$ (which exists by Lemma 3.10). Then*

$$\mathbf{E}_{\theta_1}\phi_\alpha(X) > \alpha.$$

Proof. Define $\phi'(x) := \alpha \,\forall\, x \in \mathbb{R}^n$. Let $\mathcal{T}$ be the class of all randomized hypothesis tests with significance level at most α. Since ϕ_α is UMP class $\mathcal{T}$ by Lemma 3.10, $\mathbf{E}_{\theta_1}\phi_\alpha \geq \mathbf{E}_{\theta_1}\phi' = \alpha$. It remains to eliminate the case that $\mathbf{E}_{\theta_1}\phi_\alpha(X) = \alpha$. If $\mathbf{E}_{\theta_1}\phi_\alpha(X) = \alpha$, then ϕ' is also UMP class $\mathcal{T}$. Lemma 3.10 (necessity) then implies that ϕ' is a likelihood ratio test and $\phi = \phi'$ on the set $B := \{x \in \mathbb{R}^n \colon f_{\theta_1}(x) \neq k f_{\theta_0}(x)\}$, except on a set of probability zero (with respect to both $\mathbf{P}_{\theta_0}$ and $\mathbf{P}_{\theta_1}$). Since $\alpha \in (0,1)$ we have $\phi \neq \phi'$ on B, so that $\mathbf{P}_{\theta_0}(B) = \mathbf{P}_{\theta_1}(B) = 0$. That is, $\mathbf{P}_{\theta_0}(B^c) = \mathbf{P}_{\theta_1}(B^c) = 1$. Since f_{θ_0} and f_{θ_1} are PDFs or PMFs, we must then have $k = 1$, hence $\mathbf{P}_{\theta_0} = \mathbf{P}_{\theta_1}$, a contradiction. We conclude that $\mathbf{E}_{\theta_1}\phi_\alpha(X) > \alpha$. □

3.2. **Karlin-Rubin Theorem.** The Neyman-Pearson Lemma shows that we can classify UMP tests with significance level at most α, if we want to test the alternatives between two distinct parameters. The Karlin-Rubin Theorem is another situation where a UMP test can be identified. This Theorem applies when a family of PDFs has the following property.

Definition 3.16 (**Monotone Likelihood Ratio Property**)**.** Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of PDFs where $\theta \in \Theta \subseteq \mathbb{R}$. Assume that, if $\theta_1, \theta_2 \in \mathbb{R}$ with $\theta_1 \neq \theta_2$, then $\mathbf{P}_{\theta_1} \neq \mathbf{P}_{\theta_2}$. Let $X \in \mathbb{R}^n$ have PDF f_θ. We say that $\{f_\theta \colon \theta \in \Theta\}$ has the **monotone likelihood ratio property** (MLR) if there exists a real-valued statistic $Y = t(X)$ such that, whenever $\theta_1 < \theta_2$, the ratio $f_{\theta_2}(x)/f_{\theta_1}(x)$ is a well-defined, **strictly** increasing function of $t(x)$.

Here the likelihood ratio is defined to be ∞ when $f_{\theta_2}(x) > 0$ and $f_{\theta_1}(x) = 0$. Also, the MLR property makes no assumption about $x \in \mathbb{R}^n$ such that $f_{\theta_1}(x) = f_{\theta_2}(x) = 0$.

WARNING. Many books define MLR so that the likelihood ratio is an increasing function of $t(x)$. Some other books allow the likelihood ratio to be an increasing or decreasing function of $t(x)$.

Example 3.17. Suppose we have a one-parameter exponential family of the form

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

Here $\theta \in \Theta \subseteq \mathbb{R}$. Then if $x \in \mathbb{R}^n$ satisfies $h(x) > 0$, we have

$$\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} = \exp\Big([w(\theta_2) - w(\theta_1)]t(x) - a(w(\theta_1)) + a(w(\theta_2))\Big).$$

So, if e.g. w is strictly increasing, then if $\theta_1 < \theta_2$, we have $w(\theta_2) - w(\theta_1) > 0$, so $\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}$ is a strictly increasing function of $t(x)$, i.e. this exponential family has the MLR property.

From Example 2.3, we see that a Gaussian with known variance and unknown mean μ has the MLR property with respect to μ, if we use $t(x) := (x_1 + \cdots + x_n)/n \,\forall\, x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. That is, $f_\mu \colon \mathbb{R}^n \to [0,\infty)$ is the multivariate PDF of n i.i.d. Gaussians with known variance and unknown mean $\mu \in \mathbb{R}$.

Theorem 3.18 (**Karlin-Rubin Theorem**)**.** *Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of PDFs with the MLR property, with respect to a real-valued statistic $Y = t(X)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Let*

$0 \leq \gamma \leq 1$. *Fix* $\theta_0 \in \Theta$. *Consider the hypothesis* $H_0 = \{\theta \in \Theta \colon \theta \leq \theta_0\}$ *and the hypothesis* $H_1 = \{\theta \in \Theta \colon \theta > \theta_0\}$. *Let* $c \in \mathbb{R}$. *Consider the randomized hypothesis test* $\phi \colon \mathbb{R}^n \to [0,1]$

$$\phi(x) := \begin{cases} 1 & \text{, } \textit{if } t(x) > c \\ 0 & \text{, } \textit{if } t(x) < c \\ \gamma & \text{, } \textit{if } t(x) = c. \end{cases}$$

Define $\alpha := \mathbf{E}_{\theta_0}\phi(X)$. *Let* $\mathcal{T}$ *be the class of all randomized hypothesis tests with significance level at most* α. *Then*

(i) ϕ *is UMP class* $\mathcal{T}$.

(ii) *For any* $0 < \alpha' < 1$, *there exist* $c \in \mathbb{R}$ *and* $\gamma \in [0,1]$ *such that* ϕ *is UMP class* $\mathcal{T}$ *with* $\alpha = \alpha'$.

(iii) β, *the power function of* ϕ, *is nondecreasing and strictly increasing when it takes values in* $(0,1)$.

(iv) *For any* $\theta_1 < \theta_0$, ϕ *minimizes* $\mathbf{E}_{\theta_1}\phi'$ *among all tests* ϕ' *satisfying* $\mathbf{E}_{\theta_0}\phi' = \alpha$.

Proof. We first prove (iii). Let $\theta_1 > \theta_0$ with $\theta_1 \in \Theta$ and consider the function $r \colon \mathbb{R}^n \to \mathbb{R}$ defined by

$$r(x) := \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}, \qquad \forall\, x \in \mathbb{R}^n.$$

By assumption, r is a strictly increasing function of $t(x)$. Let $k \geq 0$ such that $r(x) = k$ when $t(x) = c$. Since r is a strictly increasing function of $t(x)$, we can rewrite ϕ as

$$\phi(x) = \begin{cases} 1 & \text{, if } r(x) > k \\ 0 & \text{, if } r(x) < k \\ \gamma & \text{, if } r(x) = k. \end{cases}$$

That is, ϕ is a likelihood ratio test of the hypothesis $\{\theta = \theta_0\}$ versus $\{\theta = \theta_1\}$. Corollary 3.15 says $\beta(\theta_1) = \mathbf{E}_{\theta_1}\phi(X) > \alpha = \mathbf{E}_{\theta_0}\phi(X) = \beta(\theta_0)$, if $\mathbf{P}_{\theta_0} \neq \mathbf{P}_{\theta_1}$. (If $\mathbf{P}_{\theta_0} = \mathbf{P}_{\theta_1}$, then $\mathbf{E}_{\theta_1}\phi(X) = \mathbf{E}_{\theta_0}\phi(X) \in \{0,1\}$ since ϕ is either zero or one with probability one in this case, i.e. $\alpha \in \{0,1\}$.) Assertion (iii) follows.

We now prove (i). First, note that $\alpha = \mathbf{E}_{\theta_0}\phi(X) = \sup_{\theta \leq \theta_0} \mathbf{E}_\theta \phi(X)$ from (iii), so that ϕ is in class $\mathcal{T}$. Now let $\theta_1 > \theta_0$ with $\theta_1 \in \Theta$, and let ϕ' be a class $\mathcal{T}$ hypothesis test. By definition of $\mathcal{T}$, $\mathbf{E}_{\theta_0}\phi' \leq \sup_{\theta \leq \theta_0} \mathbf{E}_\theta \phi'(X) \leq \alpha$. So, from Lemma 3.10 (sufficiency), ϕ is UMP (in the context of that Lemma), i.e. $\mathbf{E}_{\theta_1}\phi(X) \geq \mathbf{E}_{\theta_1}\phi'(X)$. Since this inequality holds for all $\theta_1 > \theta_0$ with $\theta_1 \in \Theta$, we conclude that ϕ is UMP class $\mathcal{T}$, i.e. (i) holds.

We now prove (iv). Let $\theta_1 < \theta_0$ with $\theta_1 \in \Theta$. The MLR property now implies that $f_{\theta_0}(x)/f_{\theta_1}(x)$ is a strictly increasing function of $t(x)$. That is, there is some $k > 0$ such that

$$\phi(x) = \begin{cases} 1 & \text{, if } f_{\theta_0}(x)/f_{\theta_1}(x) > k \\ 0 & \text{, if } f_{\theta_0}(x)/f_{\theta_1}(x) < k \\ \gamma & \text{, if } f_{\theta_0}(x)/f_{\theta_1}(x) = k. \end{cases}$$

That is, ϕ is a likelihood ratio test of the hypothesis $\{\theta = \theta_1\}$ versus $\{\theta = \theta_0\}$. As in the proof of Lemma 3.10 (though with the roles of θ_0, θ_1 reversed), we have $[\mathbf{E}_{\theta_0} - k\mathbf{E}_{\theta_1}][\phi(X) - \phi'(X)] \geq 0$. So, if $\mathbf{E}_{\theta_0}\phi(X) = \mathbf{E}_{\theta_0}\phi'(X) = \alpha$, we have $\mathbf{E}_{\theta_1}\phi(X) \leq \mathbf{E}_{\theta_1}\phi'(X)$.

We now prove (ii). Define $F(k) := \mathbf{P}_{\theta_0}(t(X) \leq k)$, $\forall\, k \in \mathbb{R}$. If $\alpha' = 1$, choose $\phi := 1$. If $\alpha' = 0$, choose $\phi := 0$. Now consider the case $0 < \alpha' < 1$. Since F is monotone increasing with $\lim_{k\to-\infty} F(k) = 0$ and $\lim_{k\to\infty} F(k) = 1$, there exists $c \in \mathbb{R}$ such that $\lim_{k\to c^-} F(k) \leq 1-\alpha' \leq \lim_{k\to c^+} F(k) = F(c)$. Define then

$$\phi(x) := \begin{cases} 1 & \text{, if } t(x) > c \\ 0 & \text{, if } t(x) < c \\ \frac{F(c)-(1-\alpha')}{\mathbf{P}_{\theta_0}(t(X)=c)} & \text{, if} \, t(x) = c. \end{cases}$$

(If $\mathbf{P}_{\theta_0}(t(X) = c) = 0$, then ϕ is well-defined with $\mathbf{P}_{\theta_0}$ probability one.) Then, using the definition of ϕ,

$$\begin{aligned} \mathbf{E}_{\theta_0}\phi(X) &= \mathbf{P}_{\theta_0}(t(X) > c) + \frac{F(c)-(1-\alpha')}{\mathbf{P}_{\theta_0}(t(X)=c)}\mathbf{P}_{\theta_0}(t(X) = c) \\ &= 1 - F(c) + F(c) - (1-\alpha') = \alpha'. \end{aligned}$$

(If $\mathbf{P}_{\theta_0}(t(X) = c) = 0$, then $1 - F(c) = \alpha'$.) □

Exercise 3.19. Prove the following version of the Karlin-Rubin Theorem, with the inequalities reversed in the definition of the hypotheses.

Let $\{f_\theta\}$ be a family of PDFs with the MLR property, with respect to a real-valued statistic $Y = t(X)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Let $0 \leq \gamma \leq 1$. Fix $\theta_0 \in \Theta$. Consider the hypothesis $H_0 = \{\theta \geq \theta_0\}$ and the hypothesis $H_1 = \{\theta < \theta_0\}$. Let $c \in \mathbb{R}$. Consider the randomized hypothesis test $\phi\colon \mathbb{R}^n \to [0,1]$ defined by

$$\phi(x) := \begin{cases} 0 & \text{, if } t(x) > c \\ 1 & \text{, if } t(x) < c \\ \gamma & \text{, if } t(x) = c. \end{cases}$$

Define $\alpha := \mathbf{E}_{\theta_0}\phi(X)$. Let $\mathcal{T}$ be the class of all randomized hypothesis tests with significance level at most α.

(i) ϕ is UMP class $\mathcal{T}$.

(iii) β, the power function of ϕ, is nonincreasing and strictly decreasing when it takes values in $(0,1)$.

Example 3.20. In Example 3.17, we observed that a Gaussian with known variance and unknown mean μ has the MLR property with respect to μ, if we use $t(x) := (x_1+\cdots+x_n)/n$ $\forall\ x = (x_1,\ldots,x_n) \in \mathbb{R}^n$. That is, $f_\mu\colon \mathbb{R}^n \to [0,\infty)$ is the multivariate PDF of n i.i.d. Gaussians with known variance and unknown mean $\mu \in \mathbb{R}$. Fix $\mu_0 \in \mathbb{R}$. Consider testing the hypothesis $H_0 = \{\mu \leq \mu_0\}$ versus the hypothesis $H_1 = \{\mu > \mu_0\}$. Then Theorem 3.18 implies that the hypothesis test

$$\phi(x) := \begin{cases} 1 & \text{, if } t(x) > c \\ 0 & \text{, if } t(x) < c \\ \gamma & \text{, if } t(x) = c. \end{cases}$$

is UMP among all hypothesis test with significance level at most α, where $\alpha := \mathbf{E}_{\mu_0}\phi(X)$, and $X = (X_1,\ldots,X_n)$.

Exercise 3.21. Prove the following one-sided version of the Karlin-Rubin Theorem.

Let $\{f_\theta\}$ be a family of PDFs with the MLR property, with respect to a real-valued statistic $Y = t(X)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Let $0 \leq \gamma \leq 1$. Fix $\theta_0 \in \Theta$. Consider the hypothesis

$H_0 = \{\theta = \theta_0\}$ and the hypothesis $H_1 = \{\theta > \theta_0\}$. Let $c \in \mathbb{R}$. Consider the randomized hypothesis test $\phi\colon \mathbb{R}^n \to [0,1]$ defined by

$$\phi(x) := \begin{cases} 1 & \text{, if } t(x) > c \\ 0 & \text{, if } t(x) < c \\ \gamma & \text{, if } t(x) = c. \end{cases}$$

Define $\alpha := \mathbf{E}_{\theta_0}\phi(X)$. Let $\mathcal{T}$ be the class of all randomized hypothesis tests with significance level at most α.

Then ϕ is UMP class $\mathcal{T}$.

Exercise 3.22. Let $X_1, \ldots, X_n$ be i.i.d. random variables. Let $X = (X_1, \ldots, X_n)$. Let $\theta > 0$. Assume that X_1 is uniformly distributed in the interval $[0, \theta]$. Fix $\theta_0 > 0$. Fix $0 < \alpha < 1$. Let $\mathcal{T}$ denote the set of hypothesis tests with significance level at most α.

- Suppose we test $H_0 = \{\theta \leq \theta_0\}$ versus $H_1 = \{\theta > \theta_0\}$. Identify the set of all UMP class $\mathcal{T}$ hypothesis tests.
- Suppose we test $H_0 = \{\theta = \theta_0\}$ versus $H_1 = \{\theta \neq \theta_0\}$. Show there is a unique UMP class $\mathcal{T}$ hypothesis test in this case.

(Hint: first consider testing $\{\theta = \theta_0\}$ versus $\{\theta = \theta_1\}$ with $\theta_1 > \theta_0$, and apply the Neyman-Pearson Lemma. That is, mimic the argument of the Karlin-Rubin Theorem.) (As an aside, observe that, if you naïvely apply the Karlin-Rubin Theorem, you will not find all UMP tests, i.e. a non-strict MLR property version of the Karlin-Rubin Theorem will neglect some UMP tests.)

Exercise 3.23. This exercise demonstrates that a UMP might not always exists.

Let $X_1, \ldots, X_n$ be i.i.d. Gaussian random variables with known variance and unknown mean $\mu \in \mathbb{R}$. Fix $\mu_0 \in \mathbb{R}$. Let H_0 denote the hypothesis $\{\mu = \mu_0\}$ and let H_1 denote the hypothesis $\mu \neq \mu_0$. Fix $0 < \alpha < 1$. Let $\mathcal{T}$ denote the set of hypothesis tests with significance level at most α. Show that no UMP class $\mathcal{T}$ test exists, using the following strategy.

- Let $\mu_1 < \mu_0$. You may take as given the following fact (that follows from the Karlin-Rubin Theorem): the power at μ_1 is maximized among class $\mathcal{T}$ tests by the hypothesis test ϕ that rejects H_0 when the sample mean satisfies $\overline{X} < c$ for an appropriate choice of $c \in \mathbb{R}$. Assume for the sake of contradiction that a UMP class $\mathcal{T}$ test ϕ' exists. Then, using the necessity part of the Neyman-Pearson Lemma (i.e. consider testing $\mu = \mu_0$ versus $\mu = \mu_1$), conclude that ϕ' must have the same rejection region as ϕ (just by examining the power of the tests at μ_1.)
- Consider now a test in $\mathcal{T}$ that rejects H_0 when the sample mean satisfies $\overline{X} > c'$ for an appropriate choice of $c' \in \mathbb{R}$. Repeating the previous argument, conclude that ϕ' must reject when $\overline{X} > c'$, leading to a contradiction.

  That is, let $\mu_2 > \mu_0$. You may take as given the following fact (that follows from the Karlin-Rubin Theorem): the power at μ_2 is maximized among class $\mathcal{T}$ tests by the hypothesis test ϕ'' that rejects H_0 when the sample mean satisfies $\overline{X} > c'$ for an appropriate choice of $c' \in \mathbb{R}$. Then, using the necessity part of the Neyman-Pearson Lemma (i.e. consider testing $\mu = \mu_0$ versus $\mu = \mu_2$), conclude that ϕ' must have the same rejection region as ϕ''.

Exercise 3.24. The rejection regions C_α for UMP hypothesis tests of significance level at most $\alpha \in (0,1)$ are often nested in the sense that $C_\alpha \subseteq C_{\alpha'}$ for all $0 < \alpha < \alpha' < 1$. This exercise demonstrates an example of UMP tests where this nesting behavior does not occur.

Let $\theta_0, \theta_1 \in \mathbb{R}$ be unequal parameters. Let H_0 denote the hypothesis $\{\theta = \theta_0\}$ and let H_1 denote the hypothesis $\{\theta = \theta_1\}$. Suppose $X \in \{1,2,3\}$ is a random variable. If $\theta = \theta_0$, assume that X takes the values $1,2,3$ with probabilities $.85, .1, .05$, respectively. If $\theta = \theta_1$, assume that X takes the values $1,2,3$ with probabilities $.7, .2, .1$, respectively. Let $\mathcal{T}$ denote the set of hypothesis tests with significance level at most α.

- Let $0 < \alpha < .15$. Show that a UMP class $\mathcal{T}$ test is not unique.
- When $\alpha = .05$, show there is a unique nonrandomized hypothesis UMP class $\mathcal{T}$ test.
- When $\alpha = .1$, show there is a unique nonrandomized hypothesis UMP class $\mathcal{T}$ test.
- Show that the $\alpha = .05$ and $\alpha' = .1$ UMP nonrandomized tests from above do not have nested rejection regions.
- However, when $\alpha = .05$ and $\alpha' = .1$, there are randomized UMP tests $\phi, \phi' \colon \mathbb{R}^n \to [0,1]$ respectively, that are nested in the sense that $\phi \leq \phi'$.

3.3. Hypothesis Tests and Confidence Intervals.

Definition 3.25 (**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 3.26. 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 1.98),

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

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 3.27 (**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 nonrandomized 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 $\sup_{\theta\in\Theta_0} \beta(\theta) = \beta(\theta_0) = \mathbf{P}_{\theta_0}(X \in C(\theta_0)) \leq \alpha$. □

When we begin with a rejection region defined by a statistic, and we then obtain a confidence region via Proposition 3.27, we refer to this procedure as **inverting the test statistic**.

For example, in Example 3.26, we began with estimates for the probability that $X_1 + \cdots + X_n - n\mu$ lies outside an interval. A key property in this example is that these computed probabilities did not depend on μ. We then used some algebra to convert these probability estimates to bounds on μ, in terms of $X_1, \ldots, X_n$. For another example, suppose $X_1, \ldots, X_n$ are i.i.d. from a location family with parameter θ. Then $X_1 + \cdots + X_n - n\theta$ does not depend on θ, so we can get confidence intervals for θ if we know the distribution of $X_1 + \cdots + X_n - n\theta$, similar to what we did in Example 3.26. This procedure can often be replicated, if we can begin with a quantity whose distribution does not depend on the unknown parameter.

Definition 3.28 (**Pivotal Quantity**). Let $X \in \mathbb{R}^n$ be a random variable. Let $q\colon \mathbb{R}^n \times \Theta \to \mathbb{R}^m$. A random variable $Q := q(X, \theta)$ is called a **pivotal quantity** if the distribution of Q does not depend on θ.

WARNING. A pivotal quantity is typically **not** a statistic, since the pivotal quantity can depend on the unknown parameter. (A statistic is, by definition, a function only of the random variables. A statistic is not an explicit function of the unknown parameter.) If a pivotal quantity is a statistic, it could be called an ancillary statistic.

For the location family example mentioned above, $X_1 + \cdots + X_n - n\theta$ is a pivotal quantity, since its distribution depend on the unknown parameter $\theta \in \mathbb{R}$.

Example 3.29. Let $X_1, \ldots, X_n$ be i.i.d. exponential random variables with unknown parameter $\lambda > 0$. Suppose we want to find a confidence interval for λ. We begin by finding a pivotal quantity. It is known that $Y := X_1 + \cdots + X_n$ has a gamma distribution with parameters n and λ. As discussed in Definition 1.28, Y/λ has a gamma distribution with parameters n and 1. That is, Y/λ is a pivotal quantity, since it does not depend on λ. From Definition 1.28, we have

$$P(a \leq Y/\lambda \leq b) = \frac{1}{\Gamma(n)} \int_a^b x^{n-1} e^{-x} dx, \qquad \forall\, 0 \leq a \leq b \leq \infty.$$

We can therefore find confidence intervals for λ by writing

$$P\Big(\frac{X_1 + \cdots + X_n}{b} \leq \lambda \leq \frac{X_1 + \cdots + X_n}{a}\Big) = \frac{1}{\Gamma(n)} \int_a^b x^{n-1} e^{-x} dx, \qquad \forall\, 0 \leq a \leq b \leq \infty.$$

Exercise 3.30. Let $X_1, \ldots, X_n$ be i.i.d. Gaussian random variables with unknown mean and unknown variance.

- Find a real-valued pivotal quantity for $X = (X_1, \ldots, X_n)$.
- Using the pivotal quantity, construct a $1 - \alpha$ confidence interval for the mean μ, for any $0 < \alpha < 1$.

3.4. **Bayesian Intervals.** In Bayes estimation, the unknown parameter $\theta \in \Theta$ is regarded instead as a random variable Ψ. The distribution of Ψ reflects our prior knowledge about the probable values of Ψ. Then, given that $\Psi = \theta$, the conditional distribution of $X|\{\Psi = \theta\}$ is assumed to be $\{f_\theta \colon \theta \in \Theta\}$, where $f_\theta \colon \mathbb{R}^n \to [0, \infty)$.

Definition 3.31 (**Credible Interval, Credible Region**)**.** 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)% credible interval** for a parameter $g(\theta)$ is a random interval of the form $[u(X), v(X)]$ satisfying

$$\mathbf{P}(g(\Psi) \in [u(X), v(X)]) \geq 1 - \alpha.$$

Here $\mathbf{P}$ denotes taking a probability with respect to Ψ and X.

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

$$\mathbf{P}(g(\Psi) \in c(X)) \geq 1 - \alpha.$$

3.5. **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 3.32 (**p-Value, One-Sided, Nonrandomized**)**.** 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 3.33. 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 3.34. 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 3.35. We continue Example 3.8 and Exercise 1.94. 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.

Remark 3.33 leads to the following generalized definition of p-valued

Definition 3.36 (p-Value, Randomized). 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 $X := (X_1, \ldots, X_n)$. Consider a set of hypothesis tests $\phi_\alpha \colon \mathbb{R}^n \to [0,1]$, for any $\alpha \in [0,1]$. Assume that these tests are nested in the sense that $\phi_\alpha \leq \phi_{\alpha'}$ for all $0 \leq \alpha < \alpha' \leq 1$. The **$p$-value** for this set of hypothesis tests is the statistic

$$p(X) := \inf\{\alpha \in [0,1] \colon \phi_\alpha(X) = 1\}.$$

(If the set $\{\alpha \in [0,1] \colon \phi_\alpha(X) = 1\}$ is empty, we define $p(X) := 1$.) In the case that ϕ_α are nonrandomized, so that $\phi_\alpha = 1_{C_\alpha}$ for all $0 \leq \alpha \leq 1$, this definition becomes

$$p(X) = \inf\{\alpha \in [0,1] \colon X \in C_\alpha\}.$$

The nested property implies that $\{\alpha \in [0,1] \colon \phi_{\alpha}(X) = 1\}$ is an interval. Without the nested property, we could still define the p-value, but then it would not really be an interesting quantity to consider. (If an observation $X = x$ is rejected at a low significance level, then thinking about p-values only seems sensible when that observation is rejected at all higher significance levels.)

Exercise 3.37. 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 $X = 90$.

Exercise 3.38. Let $X_1, \ldots, X_n$ be a real-valued random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. Suppose $\Theta = \mathbb{R}$. Fix $\theta \in \mathbb{R}$. Denote $X := (X_1, \ldots, X_n)$. Consider a set of hypothesis tests $\phi_\alpha \colon \mathbb{R}^n \to [0,1]$, for any $\alpha \in [0,1]$. Assume that these tests are nested in the sense that $\phi_\alpha \leq \phi_{\alpha'}$ for all $0 \leq \alpha < \alpha' \leq 1$. Suppose we are testing the hypothesis H_0 that $\{\theta \leq \theta_0\}$ versus H_1 that $\{\theta > \theta_0\}$. Suppose also that $\{f_\theta\}$ has the monotone likelihood ratio property with respect to a statistic $Y = t(X)$ that is a continuous random variable.

- Show that the family of UMP tests with significance level at most α satisfies the nested property mentioned above (for all $\alpha \in [0,1]$).
- Show that, if $X = x$, then the p-value $p(x)$ satisfies
$$p(x) = \mathbf{P}_{\theta_0}(t(X) > t(x)).$$

3.6. **Loss Function Optimality.** As we observed in the previous section, UMP tests might not exist in fairly natural situations, such as testing $\{\theta = \theta_0\}$ versus $\{\theta \neq \theta_0\}$. To get around this issue, we can look for UMP tests in a smaller class of tests, or we could try to optimize a loss function instead of looking for a UMP test. In the former case, it is sometimes natural to search for a UMP test among all unbiased tests.

Definition 3.39. Let $\phi \colon \mathbb{R}^n \to [0,1]$ be a hypothesis test for $\{\theta \in \Theta_0\}$ versus $\{\theta \in \Theta_1\}$. Let $\beta \colon \Theta \to [0,1]$ be the power function of ϕ. We say that ϕ is **unbiased** with level $\alpha \in [0,1]$ if
$$\beta(\theta) \leq \alpha, \; \forall\, \theta \in \Theta_0, \qquad \text{and} \qquad \beta(\theta) \geq \alpha, \; \forall\, \theta \in \Theta_1.$$

Fix $0 < \alpha < 1$. If a UMP hypothesis test ϕ exists among all tests with significance level at most α, then ϕ is unbiased, since being UMP implies that $\beta(\theta) \geq \alpha$ for all $\theta \in \Theta_1$ (if we compare ϕ to the constant hypothesis test $\phi' := \alpha$). On the other hand, the class of unbiased tests with level α is smaller than the class of tests with significance level at most α. So, a finding a UMP among all unbiased tests is an optimization over a smaller class of tests, when compared with finding a UMP among all tests with a bound on their significance level.

Theorem 3.40. *Let $0 < \alpha < 1$, and let X be a random variable from a one-parameter exponential family $\{f_\theta \colon \theta \in \Theta\}$ with $w \colon \mathbb{R} \to \mathbb{R}$ (as in Definition 2.1). Assume w is continuously differentiable and strictly increasing. Let θ_0 in the interior of Θ. Let $\mathcal{T}$ denote the set of unbiased hypothesis tests with significance level at most α. Suppose we are testing the* *hypothesis $\{\theta = \theta_0\}$ versus $\{\theta \neq \theta_0\}$. Then there exists $a, b \in \mathbb{R}$ with $a \leq b$, there exists a real-valued statistic $Y = t(X)$ and a UMP class $\mathcal{T}$ test ϕ such that $\phi = 1$ when $t < a$ or $t > b$, $\phi = 0$ when $t \in (a, b)$ and such that $\mathbf{E}_{\theta_0}\phi(X)1_{t(X)<a} > 0$ and $\mathbf{E}_{\theta_0}\phi(X)1_{t(X)>b} > 0$.*

Alternatively, we could try to minimize a risk function

$$r(\theta) := \mathbf{E}_\theta \ell(\theta, \phi).$$

where ℓ is a **loss function**

$$\ell \colon \Theta \times [0, 1] \to \mathbb{R},$$

among all hypothesis tests ϕ.

4. Generalized Likelihood Ratio Tests

4.1. Generalized Likelihood Ratio Tests.

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

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

If we have data $x \in \mathbb{R}^n$, recall that we defined the function $\ell \colon \Theta \to [0, \infty)$

$$\ell(\theta) := \prod_{i=1}^n f_\theta(x_i)$$

and called it the **likelihood function**. 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 3.10 by its rejection region C'.

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

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

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

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

Definition 4.1 (**Generalized Likelihood Ratio Test**)**.** Let $k \geq 1$. The **generalized likelihood ratio test** of a hypothesis H_0 that $\{\theta \in \Theta_0\}$ is defined by the following rejection region.

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

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 4.2. If $0 < k \leq 1$ then $C = \mathbb{R}^n$. That is, all generalized likelihood ratio tests with $0 < k \leq 1$ are the same, hence our restriction to $k \geq 1$ in Definition 4.1.

Let D be the set of $x \in \mathbb{R}^n$ such that $\sup_{\theta\in\Theta_0^c} f_\theta(x) \geq \sup_{\theta\in\Theta_0} f_\theta(x)$. 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 $C \cap D^c \neq C' \cap D^c$. So, at least on the set D, the rejection regions C and C' agree.

Example 4.3. Let $X_1, \ldots, X_n$ 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 2.63, 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 3.27, the hypothesis test corresponds to confidence intervals for the sample mean. (Above we used the identity $\sum_{i=1}^n \Big[(x_i - \frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2\Big] = \sum_{i=1}^n \Big[(x_i - \mu_0 + \mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2\Big] = n(\mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2 - \frac{2}{n}\sum_{i,j=1}^n (x_i-\mu_0)(x_j-\mu_0) = n(\mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2 - 2n(\frac{1}{n}\sum_{i=1}^n(x_i-\mu_0))^2 = -n(\mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2$.)

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 2.47. 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 4.7 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 4.4. 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 4.5. 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.)

4.2. **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

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

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

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

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 $1207p_j(\lambda)$ 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 1.101, $R^{-1/2}\sqrt{1207}(\overline{Z} - p')$ 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 4.6. 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?

4.3. **Additional Comments.**

Theorem 4.7 (**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 2.69 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 2.69, 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 2.56, $\mathbf{E}_{\theta_0}\ell''(\theta_0) = -I_X(\theta_0) = -nI_{X_1}(\theta_0)$. By Theorem 2.68, $Y = Y_n$ 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 2.69, $\sqrt{n}(Y-\theta_0)$ 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 $n \to \infty$. □

5. Resampling, Bias Reduction and Estimation

The goal of bias reduction is to begin with an estimator and a random sample of fixed size n, and to find a way to reduce the bias of the estimator. We already know that conditioning as in the Rao-Blackwell Theorem 2.51 can allow us to reduce variance and maintain the bias of an estimator. Unfortunately, reducing the bias can sometimes increase the variance of the estimator. Recall that any random variable X can be written as

$$\mathbf{E}(X-\theta)^2 = \mathbf{E}(X-\mathbf{E}X+\mathbf{E}X-\theta)^2 = \mathbf{E}(X-\mathbf{E}X)^2 + (\mathbf{E}X-\theta)^2.$$

From this equality, we can intuitively assert that reducing the variance of an estimator could increase its bias, while reducing the bias of an estimator could increase its variance. This tradeoff is known as the bias-variance tradeoff.

A standard way to reduce bias is to resample from our random sample. In jackknife resampling, we consider the sample of size n with one sample removed, and then average the estimator over all n ways of removing one sample.

Another motivation for resampling methods (such as the jackknife or bootstrapping) is approximating the variance of some estimators. When the assumed probability distribution of an estimator is complicated, approximating the variance of an estimator might be complicated to do directly or require a large sample size to obtain reasonable approximations. Resampling methods allow us to approximate the variance of estimators in a way that avoids these difficulties.

Intuitively, the “extra averaging” that occurs in resampling methods leads to more accurate estimates.

5.1. Jackknife Resampling.

Definition 5.1. Let $\Theta\subseteq\mathbb{R}$. Let $X_1,X_2,\ldots\colon\Omega\to\mathbb{R}^d$ be i.i.d random variables so that X_1 has distribution $f_\theta\colon\mathbb{R}^d\to[0,\infty)$, $\theta\in\Theta$. Let $Y_1,Y_2,\ldots$ be a sequence of estimators for θ so that for any $n\geq 1$, $Y_n=t_n(X_1,\ldots,X_n)$ for some $t_n\colon\mathbb{R}^{nd}\to\Theta$. For any $n\geq 1$, define the **jackknife estimator** of Y_n to be

$$Z_n := nY_n - \frac{n-1}{n}\sum_{i=1}^{n} t_{n-1}(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n).$$

Define also the **jackknife estimator for the bias** of Y_n to be

$$Y_n - Z_n.$$

The jackknife estimator reduces the bias of the original estimator, as we now show.

Proposition 5.2. *Assume that $Y_1,Y_2,\ldots$ are asymptotically unbiased, so that there exists $a,b\in\mathbb{R}$ such that*

$$\mathbf{E}Y_n = \theta + a/n + b/n^2 + O(1/n^3), \qquad \forall\, n\geq 1. \qquad (*)$$

Then

$$\mathbf{E}Z_n = \theta + O(1/n^2).$$

And if $b=0$ and the $O(1/n^3)$ term is zero in $(*)$, then Z_n is unbiased.

Proof. Let $n \geq 1$. Then

$$\begin{aligned}
\mathbf{E}Z_n &\overset{(*)}{=} n\theta + a + \frac{b}{n} + O(1/n^2) - \frac{n-1}{n}\sum_{i=1}^{n}\mathbf{E}t_{n-1}(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)\\
&\overset{(*)}{=} n\theta + a + \frac{b}{n} + O(1/n^2) - \frac{n-1}{n}\sum_{i=1}^{n}(\theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O(1/n^3))\\
&= \theta + \frac{b}{n} - \frac{b}{n-1} + O(1/n^2) = \theta + O(1/n^2).
\end{aligned}$$

□

Example 5.3. The jackknife estimator of the sample mean is the sample mean.

$$\begin{aligned}
&nY_n - \frac{n-1}{n}\sum_{i=1}^{n} t_{n-1}(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)\\
&\quad= \sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}(X_1+\cdots+X_{i-1}+X_{i+1}+\cdots+X_n)\\
&\quad= \sum_{i=1}^{n} X_i - \frac{n-1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \forall\, n \geq 1.
\end{aligned}$$

Example 5.4. Let $X_1,\ldots,X_n$ be i.i.d. Bernoulli random variables with parameter $0<\theta<1$. The MLE for θ is the sample mean, so by the Functional Equivariance Property of the MLE, Proposition 2.65, the MLE for θ^2 is

$$Y_n := \Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)^2, \qquad \forall\, n \geq 1.$$

This estimator is biased, since

$$\mathbf{E}Y_n = \frac{1}{n^2}\Big(n\theta + n(n-1)\theta^2\Big) = \theta^2 + \frac{1}{n}(\theta - \theta^2), \qquad \forall\, n \geq 1.$$

By Proposition 5.2, the jackknife estimator

$$Z_n := n\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)^2 - \frac{n-1}{n}\sum_{i=1}^{n}\Big(\frac{1}{n-1}\sum_{j\in\{1,\ldots,n\}\colon\, j\neq i} X_j\Big)^2, \qquad \forall\, n \geq 1.$$

is an unbiased estimator of θ^2.

5.2. **Jackknife Variance Estimator.** We begin by rewriting the jackknife estimator from Definition 5.1 as

$$\begin{aligned}
Z_n &:= nY_n - \frac{n-1}{n}\sum_{i=1}^{n} t_{n-1}(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)\\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(nY_n - (n-1)t_{n-1}(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)\Big) =: \frac{1}{n}\sum_{i=1}^{n} Y_{n,i}.
\end{aligned}$$

Using the heuristic assumption that the terms in the sum behave as i.i.d. random variables with variance $\mathrm{var}(\sqrt{n}Y_n)$, we obtain the following estimator of $\mathrm{var}(Y_n)$, which can be considered a sample variance of $Y_{n,1}/\sqrt{n},\ldots,Y_{n,n}/\sqrt{n}$.

Definition 5.5. The **jackknife variance estimator** of Y_n is

$$V_n := \frac{1}{n-1}\sum_{i=1}^{n}\Big(\frac{Y_{n,i}}{\sqrt{n}} - \frac{1}{n}\sum_{j=1}^{n}\frac{Y_{n,j}}{\sqrt{n}}\Big)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\Big(Y_{n,i} - \frac{1}{n}\sum_{j=1}^{n}Y_{n,j}\Big)^2$$
$$= \frac{n-1}{n}\sum_{i=1}^{n}\Big(t_{n-1}(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n) - \frac{1}{n}\sum_{j=1}^{n}t_{n-1}(X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_n)\Big)^2.$$

Example 5.6. The jackknife variance estimator of the sample mean is a multiple of the sample variance itself.

$$\begin{aligned}
&\frac{n-1}{n}\sum_{i=1}^{n}\Big(\frac{1}{n-1}\sum_{j\neq i}X_j - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{n-1}\sum_{k\neq j}X_k\Big)^2\\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\Big(\sum_{j\neq i}X_j - \frac{1}{n}\sum_{j=1}^{n}\sum_{k\neq j}X_k\Big)^2\\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\Big(-\frac{n-1}{n}X_i + (1-(n-1)/n)\sum_{j\neq i}X_j\Big)^2\\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\Big(-\frac{n-1}{n}X_i + (1/n)\sum_{j\neq i}X_j\Big)^2\\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\Big(-X_i + \frac{1}{n}\sum_{j=1}^{n}X_j\Big)^2.
\end{aligned}$$

Example 5.7. Let $X_1,\ldots,X_n$ be i.i.d. Bernoulli random variables with parameter $0<\theta<1$. The MLE for θ is the sample mean, so by the Functional Equivariance Property of the MLE, Proposition 2.65, the MLE for θ^2 is

$$Y_n := \Big(\frac{1}{n}\sum_{i=1}^{n}X_i\Big)^2, \qquad \forall\, n\geq 1.$$

The jackknife estimator for the variance of Y_n is

$$V_n := \frac{n-1}{n}\sum_{i=1}^{n}\Big[\Big(\frac{1}{n-1}\sum_{k\neq i}X_k\Big)^2 - \frac{1}{n}\sum_{j=1}^{n}\Big(\frac{1}{n-1}\sum_{k\neq j}X_k\Big)^2\Big]^2, \qquad \forall\, n\geq 1.$$

Despite its heuristic definition, the jackknife variance estimator does estimate the variance of Y_n as $n\to\infty$, as the following Theorem demonstrates.

Theorem 5.8 (Consistency of Jackknife Variance Estimator). *Let $X_1,X_2,\ldots\colon\Omega\to\mathbb{R}^d$ be i.i.d random variables. Let $Y_1,Y_2,\ldots$ be a sequence of real-valued estimators for θ so that for any $n\geq 1$, $Y_n = t(\overline{X}_n)$ for some $t\colon\mathbb{R}^d\to\mathbb{R}$. Assume that $\mu := \mathbf{E}X_1\in\mathbb{R}^d$. Let $\Sigma\in\mathbb{R}^{d^2}$ denote the covariance matrix of X_1 (so that $\Sigma_{ij} := \mathrm{cov}(X_{1i},X_{1j})$ is finite and exists $\forall\ 1\leq i,j\leq d$.) Assume that $\Sigma_{ij}\in\mathbb{R}$ for all $1\leq i,j\leq d$. Assume that ∇t exists and is continuous in a neighborhood of μ, and $\nabla t(\mu)\neq 0$. Then $V_1,V_2,\ldots$ is strongly consistent, in the sense that*

$$\frac{V_n}{\frac{1}{n}[\nabla t(\mu)]^T\Sigma[\nabla t(\mu)]}$$

converges almost surely to 1 *as* $n \to \infty$.

Recall that almost sure convergence implies convergence in probability by Exercise 1.85, so strong consistency as stated here implies consistency (see Definition 2.44).

Recall also that the multivariate Central Limit Theorem 1.101 implies that

$$\frac{Y_n - t(\mu)}{\sqrt{\frac{1}{n}[\nabla t(\mu)]^T \Sigma [\nabla t(\mu)]}}$$

converges in distribution to a standard Gaussian random vector in $\mathbb{R}^d$ as $n \to \infty$, explaining why we consider $\frac{1}{n}[\nabla t(\mu)]^T \Sigma [\nabla t(\mu)]$ to be the asymptotic variance of Y_n as $n \to \infty$.

Proof. Denote $\overline{X}_{n,i} := (X_1 + \cdots + X_{i-1} + X_{i+1} + \cdots + X_n)/(n-1)$ for all $1 \leq i \leq n$. Denote $T_{n,i} := t_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) = t(\overline{X}_{n,i})$ and $T_n := t_n(X_1, \ldots, X_n) = t(\overline{X}_n)$. Then

$$T_{n,i} - T_n = t(\overline{X}_{n,i}) - t(\overline{X}_n) = [\nabla t(\xi_{n,i})]^T (\overline{X}_{n,i} - \overline{X}_n) = [\nabla t(\overline{X}_n)]^T (\overline{X}_{n,i} - \overline{X}_n) + R_{n,i}. \qquad (*)$$

Here $\xi_{n,i} \in \mathbb{R}^d$ lies on the straight line between $\overline{X}_{n,i}$ and $\overline{X}_n$, and $R_{n,i} := [\nabla t(\xi_{n,i}) - \nabla t(\overline{X}_n)]^T (\overline{X}_{n,i} - \overline{X}_n)$. Denote $\overline{R}_n := \frac{1}{n}\sum_{i=1}^n R_{n,i}$. From the Mean Value Theorem,

$$\begin{aligned} V_n &= \frac{n-1}{n}\sum_{i=1}^n \Big(T_{n,i} - \frac{1}{n}\sum_{j=1}^n T_{n,j}\Big)^2 = \frac{n-1}{n}\sum_{i=1}^n \Big(T_{n,i} - T_n - \frac{1}{n}\sum_{j=1}^n [T_{n,j} - T_n]\Big)^2 \\ &\overset{(*)}{=} \frac{n-1}{n}\sum_{i=1}^n \Big([\nabla t(\overline{X}_n)]^T (\overline{X}_{n,i} - \overline{X}_n) + R_{n,i} - \frac{1}{n}\sum_{j=1}^n [T_{n,j} - T_n]\Big)^2 \end{aligned}$$

The last term simplifies since

$$\sum_{i=1}^n [T_{n,i} - T_n] \overset{(*)}{=} [\nabla t(\overline{X}_n)]^T \sum_{i=1}^n (\overline{X}_{n,i} - \overline{X}_n) + \sum_{i=1}^n R_{n,i} = n\overline{R}_n,$$

using

$$(\overline{X}_{n,i} - \overline{X}_n) = (n-1)^{-1}(\overline{X}_n - X_i), \qquad \forall\, 1 \leq i \leq n. \qquad (**)$$

Therefore

$$\begin{aligned} V_n &= \frac{n-1}{n}\sum_{i=1}^n \Big([\nabla t(\overline{X}_n)]^T (\overline{X}_{n,i} - \overline{X}_n) + R_{n,i} - \overline{R}_n\Big)^2 \\ &= \frac{n-1}{n}[\nabla t(\overline{X}_n)]^T \Big(\sum_{i=1}^n (\overline{X}_{n,i} - \overline{X}_n)(\overline{X}_{n,i} - \overline{X}_n)^T\Big) \nabla t(\overline{X}_n) \\ &\quad + \frac{n-1}{n}\sum_{i=1}^n [\nabla t(\overline{X}_n)]^T (\overline{X}_{n,i} - \overline{X}_n)[R_{n,i} - \overline{R}_n] + \frac{n-1}{n}\sum_{i=1}^n [R_{n,i} - \overline{R}_n]^2 \end{aligned}$$

We begin with the first term. From $(**)$, the first term can be written as

$$\frac{1}{n(n-1)}[\nabla t(\overline{X}_n)]^T \Big(\sum_{i=1}^n (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T\Big) \nabla t(\overline{X}_n)$$

The $\nabla t(\overline{X}_n)$ term converges almost surely to $\nabla t(\mu)$ as $n \to \infty$ by the Strong Law of Large numbers (Theorem 1.88), and continuity of ∇t at μ. Multiplying by n and applying the Strong Law of Large numbers again to the sum over i, the entire term converges almost surely to $[\nabla t(\mu)]^T \Sigma [\nabla t(\mu)]$ as $n \to \infty$. (More specifically, apply the Law of Large numbers first to $\sum_{i=1}^n X_i^T X_i$, and so on.)

We will show that the last term multiplied by n converges to zero almost surely, and this completes the proof since then the middle term converges to zero by the Cauchy-Schwarz inequality. To control the last term we write (first using $\mathbf{E}(Z - \mathbf{E}Z)^2 \leq \mathbf{E}Z^2$)

$$(n-1)\sum_{i=1}^n [R_{n,i} - \overline{R}_n]^2 \leq (n-1)\sum_{i=1}^n R_{n,i}^2 \leq \max_{1 \leq j \leq n} \|\nabla t(\xi_{n,j}) - \nabla t(\overline{X}_n)\|^2 \cdot (n-1)\sum_{i=1}^n \|\overline{X}_{n,i} - \overline{X}_n\|^2.$$

From $(**)$, the right term satisfies

$$(n-1)\sum_{i=1}^n \|\overline{X}_{n,i} - \overline{X}_n\|^2 = \frac{1}{n-1}\sum_{i=1}^n \|X_i - \overline{X}_n\|^2,$$

which converges a.s. to the trace of Σ as $n \to \infty$, by the Strong Law of Large Numbers. Meanwhile, by definition of $\xi_{n,i}$, we have $\|\xi_{n,i} - \overline{X}_n\|$ going to zero a.s. as $n \to \infty$ for each fixed $1 \leq i \leq n$, but we need this convergence to occur uniformly over all $1 \leq i \leq n$, since will take a maximum of these terms. To this end, observe that

$$\|\xi_{n,i} - \overline{X}_n\|^2 \leq \|\overline{X}_{n,i} - \overline{X}_n\|^2 \overset{(**)}{=} \frac{1}{(n-1)^2}\|X_i - \overline{X}_n\|^2 \leq \frac{1}{(n-1)^2}\|X_i\|^2, \qquad \forall\, 1 \leq i \leq n.$$

This leads to a uniform bound in i since, for any $s > 0$,

$$\begin{aligned}\mathbf{P}(\max_{1 \leq i \leq n} \|X_i/n\| > s) &= 1 - \mathbf{P}(\max_{1 \leq i \leq n} \|X_i/n\| \leq s) = 1 - [\mathbf{P}(\|X_1\| \leq sn)]^n \\ &= 1 - e^{n\log(1 - \mathbf{P}(\|X_1\| > sn))} = 1 - e^{-n[\mathbf{P}(\|X_1\| > sn) + o(\mathbf{P}(\|X_1\| > sn))]}.\end{aligned}$$

Since $\mathbf{E}\|X_1\| = \int_0^\infty \mathbf{P}(\|X_1\| > s)ds < \infty$, $\lim_{n\to\infty} n\mathbf{P}(\|X_1\| > n) = 0$, so

$$\lim_{n\to\infty} \mathbf{P}(\max_{1 \leq i \leq n} \|X_i/n\| > s) = 0, \qquad \forall\, s > 0.$$

(If $\lim_{n\to\infty} n\mathbf{P}(\|X_1\| > n) \neq 0$, then there are $n_1, n_2, \ldots \geq 5$ with $n_i\mathbf{P}(\|X_1\| > n_i) > \varepsilon > 0$ for all $i \geq 1$, so that for any $0 < s < 1$, $(n_i - s)\mathbf{P}(\|X_1\| > n_i - s) \geq (n_i - s)\mathbf{P}(\|X_1\| > n_i) \geq \varepsilon/2$ for all $i \geq 1$, so that $\mathbf{E}\|X_1\| = \int_0^\infty \mathbf{P}(\|X_1\| > s)ds = \infty$, a contradiction.)

Consequently, $\max_{1 \leq i \leq n} \|\xi_{n,i} - \overline{X}_n\|^2$ converges to zero a.s. as $n \to \infty$, so by continuity of ∇t at μ, we conclude that $\max_{1 \leq j \leq n} \|\nabla t(\xi_{n,j}) - \nabla t(\overline{X}_n)\|^2$ also converges to zero a.s. as $n \to \infty$, as desired. □

5.3. **Bootstrapping.**

Definition 5.9. Let $X_1, \ldots, X_n$ be a random sample of size n. Let $m \geq 1$. We define the **bootstrap sample** $W_1, \ldots, W_m$ as follows. Given $X_1, \ldots, X_n$, let $W_1, \ldots, W_m$ be a random sample of size m uniformly distributed in the values $\{X_1, \ldots, X_n\}$.

We typically take m significantly larger than n.

For example, if we are given a sample of the form $\{3, 3, 5, 6\}$, then W_1 has probability $1/2$ of taking the value 3.

Remark 5.10. Note that $W_1, \ldots, W_m$ are conditionally independent, by their definition. Although the original sample consists of independent random variables, the bootstrap sample does not. The easiest way to see this is to show that the covariance of W_1 and W_2 is nonzero. Indeed, using the conditional independence, we have

$$\begin{aligned}\mathbf{E}W_1W_2 &= \mathbf{E}\Big[\mathbf{E}(W_1W_2|X_1,\ldots,X_n)\Big] = \mathbf{E}\Big[\mathbf{E}(W_1|X_1,\ldots,X_n)\cdot\mathbf{E}(W_2|X_1,\ldots,X_n)\Big]\\ &= \mathbf{E}\Big[\Big(\mathbf{E}(W_1|X_1,\ldots,X_n)\Big)^2\Big] = \mathbf{E}\overline{X}^2.\end{aligned}$$

Meanwhile

$$\mathbf{E}(\overline{W}|X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^n\Big(\frac{1}{n}\sum_{j=1}^n X_j\Big) = \overline{X}. \qquad (\ddagger)$$

So, the covariance of W_1 and W_2 is

$$\mathbf{E}(W_1-\mathbf{E}W_1)(W_2-\mathbf{E}W_2) = \mathbf{E}W_1W_2-(\mathbf{E}W_1)(\mathbf{E}W_2) = \mathbf{E}\overline{X}^2-(\mathbf{E}\overline{X})^2 = \mathrm{Var}\overline{X} = \frac{\mathrm{Var}(X_1)}{n}.$$

So, if X_1 is nonconstant, this covariance is nonzero.

Definition 5.11. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}^n$ be i.i.d random variables. Let $Y_1, Y_2, \ldots$ be a sequence of real-valued estimators so that for any $n \geq 1$, $Y_n = t_n(X_1,\ldots,X_n)$ for some $t_n \colon \mathbb{R}^{n^2} \to \mathbb{R}$. For any $n \geq 1$, define the **bootstrap variance estimator** of Y_n as

$$\mathrm{Var}(t_n(W_1,\ldots,W_n)\,|\,X_1,\ldots,X_n),$$

where $W_1, \ldots, W_n$ is the bootstrap sample of $X_1, \ldots, X_n$.

Recall that $\mathrm{Var}(W|X) := \mathbf{E}((W-\mathbf{E}(W|X))^2|X)$. The bootstrap variance estimator is sometimes called a **nonparametric bootstrap** variance estimator, since we did not use any assumptions about unknown parameters.

Example 5.12. Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean $\mu \in \mathbb{R}$. Denote $X := (X_1, \ldots, X_n)$. Let Y_n denote the sample mean of $X_1, \ldots, X_n$. Then the bootstrap variance estimator of Y_n is

$$\begin{aligned}&\mathrm{Var}(t_n(W_1,\ldots,W_n)\,|\,X) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^n W_i\,\Big|\,X\Big) = \mathbf{E}\Big(\Big[\frac{1}{n}\sum_{i=1}^n W_i - \mathbf{E}\Big(\frac{1}{n}\sum_{i=1}^n W_i\,|\,X\Big)\Big]^2\,\Big|\,X\Big)\\ &= \mathbf{E}\Big(\Big[\frac{1}{n}\sum_{i=1}^n W_i - \frac{1}{n}\sum_{i=1}^n X_i\Big]^2\,\Big|\,X\Big) = \mathbf{E}\Big(\Big[\frac{1}{n}\sum_{i=1}^n W_i\Big]^2\,\Big|\,X\Big) - \Big(\frac{1}{n}\sum_{i=1}^n X_i\Big)^2\\ &= \mathbf{E}\Big(\frac{1}{n}W_1^2 + \frac{n^2-n}{n^2}W_1W_2\,\Big|\,X\Big) - \Big(\frac{1}{n}\sum_{i=1}^n X_i\Big)^2\\ &= \frac{1}{n^2}\sum_{i=1}^n X_i^2 + \frac{n^2-n}{n^2}\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big)^2 - \Big(\frac{1}{n}\sum_{i=1}^n X_i\Big)^2\\ &= \frac{1}{n^2}\sum_{i=1}^n X_i^2 - \frac{1}{n}\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big)^2 = \frac{1}{n^2}\sum_{i=1}^n\Big(X_i - \frac{1}{n}\sum_{j=1}^n X_j\Big)^2.\end{aligned}$$

In practice, the bootstrap variance estimator can be difficult to use directly, since the conditional variance expression could become complicated. A more practical approximate variance estimator is then

$$V_m := \frac{1}{m}\sum_{i=1}^{m}\Big(t_n(W_{1i},\ldots,W_{ni}) - \frac{1}{m}\sum_{j=1}^{m}t_n(W_{1j},\ldots,W_{nj})\Big)^2.$$

Here $(W_{1i},\ldots,W_{ni})$ are (conditionally) independent bootstrap samples from $X_1,\ldots,X_n$, for each $1\leq i\leq m$. From this independence, the Strong Law of Large numbers (Theorem 1.88) implies (under a finite second moment assumption) that V_m converges almost surely to the bootstrap variance estimator as $m\to\infty$.

Similarly, we could estimate the bias of the estimator by

$$\frac{1}{m}\sum_{i=1}^{m}\Big(t_n(W_{1i},\ldots,W_{ni}) - t_n(X_1,\ldots,X_n)\Big).$$

Theorem 5.13 (**Consistency of Bootstrap**). *Let $X_1,X_2,\ldots\colon\Omega\to\mathbb{R}^d$ be i.i.d random variables. Assume $\mathbf{E}\left\|X_1\right\|^2<\infty$. Let $Y_1,Y_2,\ldots$ be a sequence of real-valued estimators so that for any $n\geq 1$, $Y_n=t(\overline{X}_n)$ for some $t\colon\mathbb{R}^d\to\mathbb{R}$. Let $Z_1,Z_2,\ldots$ be the corresponding bootstrap estimators, so that $Z_n=t(\overline{W}_n)$. Assume that $\mu:=\mathbf{E}X_1\in\mathbb{R}^d$. Assume that ∇t exists and is continuous in a neighborhood of μ, and $\nabla t(\mu)\neq 0$. Then $Z_1,Z_2,\ldots$ is strongly consistent, in the sense that*

$$\sup_{a\in\mathbb{R}}\Big|\mathbf{P}(Y_n\leq a)-\mathbf{P}(Z_n\leq a\,|\,(X_1,\ldots,X_n))\Big|.$$

converges to zero almost surely as $n\to\infty$.

5.4. **Bootstrap Confidence Intervals.** In Example 3.17, we constructed confidence intervals for a real valued Gaussian with known variance σ^2 and unknown mean μ using the pivotal quantity $(\overline{X}_n-\mu)/(\sigma/\sqrt{n})$, since this quantity is a mean zero variance one Gaussian. Let $F\colon\mathbb{R}\to[0,1]$ denote the CDF of a mean zero variance one Gaussian. For any $a\geq b$, we then had

$$\mathbf{P}(\overline{X}_n-a\sigma/\sqrt{n}\leq\mu\leq\overline{X}_n-b\sigma/\sqrt{n})=F(-b)-F(-a).$$

Fix $\alpha\in(0,1/2)$, and let $a:=-F^{-1}(\alpha)=F^{-1}(1-\alpha)$, $b:=-F^{-1}(1-\alpha)=F^{-1}(\alpha)$ to get

$$\mathbf{P}(\overline{X}_n-F^{-1}(1-\alpha)\sigma/\sqrt{n}\leq\mu\leq\overline{X}_n-F^{-1}(\alpha)\sigma/\sqrt{n})=(1-\alpha)-\alpha=1-2\alpha.$$

So, in this Gaussian setting, we have a $1-2\alpha$ confidence interval for μ of the form

$$[\overline{X}_n-F^{-1}(1-\alpha)\sigma/\sqrt{n},\ \overline{X}_n-F^{-1}(\alpha)\sigma/\sqrt{n}].$$

We note in passing (with Theorem 5.14 below in mind) that

$$\mathbf{P}(\overline{X}_n-a\sigma/\sqrt{n}\leq\mu)=\mathbf{P}(\overline{X}_n-F^{-1}(1-\alpha)\sigma/\sqrt{n}\leq\mu)=F(F^{-1}(1-\alpha))=1-\alpha.$$

We can mimic this approach for bootstrap estimators. Let $X_1,\ldots,X_n\in\mathbb{R}^d$ be i.i.d. random variables. Let $W_1,\ldots,W_m$ be a bootstrap sample from $X_1,\ldots,X_n$. Let $T_n\in\mathbb{R}$ be an estimator of θ, and let S_n^2 be an estimator of the variance of T_n. Let $T_{n,b}$ be the bootstrap version of T_n, and let $S_{n,b}^2$ be the bootstrap version of S_n^2. Let $F_b\colon\mathbb{R}\to[0,1]$ denote the CDF of $(T_{n,b}-T_n)/S_{n,b}$.

$$F_b(u):=\mathbf{P}\Big((T_{n,b}-T_n)/S_{n,b}\leq u\,\Big|\,T_n\Big),\qquad\forall\,u\in\mathbb{R}.$$

Then

$$[T_n - F_b^{-1}(1-\alpha)S_n,\ T_n - F_b^{-1}(\alpha)S_n]$$

is an approximate $1-2\alpha$ confidence interval for θ.

Theorem 5.14 (**Bootstrap Confidence Set Consistency**). *Let $X_1,\ldots,X_n \in \mathbb{R}^d$ be i.i.d. random variables such that X_1 has distribution $\{f_\theta\}$ where $\theta \in \mathbb{R}$ is unknown. Let $Y_1, Y_2, \ldots$ be a sequence of real-valued estimators for θ. Let $Z_1, Z_2, \ldots$ be the corresponding bootstrap estimators. Assume that*

$$\sup_{a\in\mathbb{R}}\Big|\mathbf{P}(Y_n \leq a) - \mathbf{P}(Z_n \leq a \,|\, (X_1,\ldots,X_n))\Big|.$$

converges to zero almost surely as $n \to \infty$. Then

$$\lim_{n\to\infty}\mathbf{P}(T_n - F_b^{-1}(1-\alpha)S_n \leq \theta) = 1-\alpha.$$

Exercise 5.15. Let $X_1,\ldots,X_n$ be i.i.d. random variables. Let $0 < \alpha < 1/2$. Define the α-trimmed sample mean to be

$$\overline{X}_n^{(\alpha)} := \frac{1}{n - 2\lfloor n\alpha\rfloor}\sum_{i=\lfloor n\alpha\rfloor+1}^{n-\lfloor n\alpha\rfloor} X_{(i)}.$$

For any $w = (w_1,\ldots,w_n) \in \{1,\ldots,n\}^n$, define the Winsorized sample mean to be

$$\overline{X}_n^{(w)} := \frac{1}{n}\sum_{i=1}^n X_{(w_i)}.$$

- Show that the jackknife estimator of $\overline{X}_n^{(\alpha)}$ is

$$\frac{1}{1-2\alpha}(\overline{X}_n^{(w)} - 2\alpha\overline{X}_n^{(\alpha)}),$$

  for some vector w.
- Show that the jackknife variance estimator of $\overline{X}_n^{(\alpha)}$ is

$$\frac{1}{n(n-1)(1-2\alpha)^2}\sum_{i=1}^n(\overline{X}_{(w_i)} - \overline{X}_n^{(w)})^2,$$

  for some vector w.

Exercise 5.16. Let X_1, X_2, X_3 be i.i.d. continuous random variables such that X_1 has PDF $\{f_\theta \colon \theta \in \Theta\}$. Let W_1, W_2, W_3 be a bootstrap sample from X_1, X_2, X_3. Let Y denote the sample median of X_1, X_2, X_3. (That is, Y is the middle value among X_1, X_2, X_3, which is unique with probability one since the random variables are continuous.)

- Describe the distribution of $(W_{(1)}, W_{(2)}, W_{(3)})$.
- Describe the bootstrap estimator of Y.
- Describe the bootstrap estimator of the variance of Y.

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 $X_1,\ldots,X_n$ is independent of $Y_1,\ldots,Y_m$, 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 $\sigma_X^2+\sigma_Y^2$. 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 $X_1,\ldots,X_n$ 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 $Y_1,\ldots,Y_m$ 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 $X_1,\ldots,X_n$ is independent of $Y_1,\ldots,Y_m$, 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.

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.

In the case that A^TA is invertible, we can multiply both sides of Y by $(A^TA)^{-1}A^T$ to get

$$\beta=(A^TA)^{-1}A^T(Y-\varepsilon).$$

If ε is a mean zero vector, then the estimator

$$Z:=(A^TA)^{-1}A^TY$$

is unbiased, i.e. its expected value is β. Its covariance matrix $\mathrm{Cov}(Z)$ is

$$(A^TA)^{-1}A^T\mathrm{Cov}(Y)A(A^TA)^{-1}=(A^TA)^{-1}A^T\mathrm{Cov}(\varepsilon)A(A^TA)^{-1}.$$

When ε is a vector of i.i.d. random variables each with variance σ^2, this reduces to

$$\sigma^2(A^TA)^{-1}.$$

Exercise 7.1. Under the above assumptions, show that the estimator

$$\Big(\frac{1}{n-m}\sum_{i=1}^{n}(Y_i-(AZ)_i)^2\Big)(A^TA)^{-1}$$

is an unbiased estimator of the covariance matrix of $Z:=(A^TA)^{-1}A^TY$.

When A is a rank two matrix corresponding to linear regression (see Example 7.3), the estimator $(A^T A)^{-1}A^T Y$ has a simpler form since

$$A^T A = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}.$$

$$(A^T A)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

$$(A^T A)^{-1}A^T Y = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 \sum_{j=1}^n Y_j - \sum_{i=1}^n x_i \sum_{j=1}^n Y_j x_j \\ -\sum_{i=1}^n x_i \sum_{j=1}^n Y_j + n\sum_{i=1}^n Y_i x_i \end{pmatrix}.$$

Theorem 7.9 demonstrates that this estimator has minimal variance among unbiased linear estimators of β.

Suppose we want to estimate some real parameter $g(\beta)$ where $g\colon \mathbb{R}^m \to \mathbb{R}$. We could estimate $g(\beta)$ with $g(Z)$. A jackknife estimator of the variance of $g(Z)$ would be

$$\frac{n-1}{n}\sum_{i=1}^n \Big(g(Z_i) - \frac{1}{n}\sum_{j=1}^n g(Z_j)\Big)^2,$$

where $Z_j := (A_j^T A_j)^{-1}A_j^T Y_j$, and A_j denotes A with the j^{th} row removed, for each $1 \leq j \leq n$ (and similarly for Y_j). This estimator is consistent under some reasonable assumptions, but the jackknife mean bias estimator

$$\frac{n-1}{n}\sum_{i=1}^n \Big(g(Z_i) - g(Z)\Big),$$

is not consistent in general.

Example 7.2 (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 7.3 (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

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

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.

More generally, Let $\beta_0,\ldots,\beta_p\in\mathbb{R}$ be unknown. Let $\{x_{ij}\}_{1\leq i\leq n,1\leq j\leq p}\in\mathbb{R}$ 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

$$Y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}+\varepsilon_i,\qquad\forall\,1\leq i\leq n.$$

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 $1\leq j\leq p$, 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

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

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

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

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

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 pivotal quantity. 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 pivotal quantity.

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 7.4. *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 $1 \leq i < j \leq p$, and set $c_i = 1$, $c_j = -1$ and $c_k = 0$ for all other $k \in \{1, \ldots, p\}$. 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 $c_1, \ldots, c_p \in \mathbb{R}$ with $\sum_{i=1}^p c_i = 0$, we have

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

Proposition 7.5. *Define*

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

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 2.18.)*

Proof. Apply Lemma 7.7 with $a_i = n_i^{-1}$, $b_i := \overline{Y}_i - \beta_i$ $\forall$ $1\leq i\leq p$, noting that

$$\begin{aligned}\frac{\Big(\sum_{j=1}^p c_j\overline{Y}_j - \sum_{j=1}^p c_j\beta_j\Big)^2}{\sum_{j=1}^p \frac{c_j^2}{n_j}} &= \frac{t^2}{\sum_{i=1}^p a_ic_i^2} = \sum_{\ell=1}^p a_\ell^{-1}\Big(b_\ell - \frac{\sum_{j=1}^p b_ja_j^{-1}}{\sum_{k=1}^p a_k^{-1}}\Big)^2\\ &= \sum_{\ell=1}^p n_\ell\Big(b_\ell - \frac{\sum_{j=1}^p b_jn_j}{\sum_{k=1}^p n_k}\Big)^2 = \sum_{j=1}^p n_j((\overline{Y_j}-\overline{Y})-(\beta_i-\overline{\beta}))^2\end{aligned}$$

Finally, (a generalization of) Proposition 2.15 implies that the numerator and denominator of F are independent, and Exercise 2.18 completes the proof. □

Remark 7.6. 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 7.7. *Let* $a_1,\ldots,a_n>0$, *let* $b_1,\ldots,b_n\in\mathbb{R}$ *and let* $t\neq 0$. *Suppose we minimize*

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

subject to the constraints

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

Then the minimum value of this problem occurs when

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

$$c_i = \frac{ta_i^{-1}\Big(b_i - \frac{\sum_{j=1}^n b_ja_j^{-1}}{\sum_{k=1}^n a_k^{-1}}\Big)}{\sum_{\ell=1}^n a_\ell^{-1}\Big(b_\ell - \frac{\sum_{j=1}^n b_ja_j^{-1}}{\sum_{k=1}^n a_k^{-1}}\Big)^2}, \qquad \forall 1\leq i\leq n.$$

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

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

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,

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

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,

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

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 □

7.3. **Linear Regression.**

Exercise 7.8. 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 7.3, 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 7.9. *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 $c_1, \ldots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^{n} c_i Y_i$ is an unbiased estimator of β_1. Suppose

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

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

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

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

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

Suppose we minimize this quantity subject to the constraint $(*)$. Lemma 7.7 with $t = 1$, $b_i = x_i$ and $a_i = 1$ for all i implies that this minimum occurs when

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

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

Step 2. Since

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

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

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

Suppose we minimize this quantity subject to the constraint $(**)$. Lemma 7.7 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

$$c_i' = \frac{ta_i^{-1}\left(b_i - \frac{\sum_{j=1}^n b_j a_j^{-1}}{\sum_{k=1}^n a_k^{-1}}\right)}{\sum_{\ell=1}^n a_\ell^{-1}\left(b_\ell - \frac{\sum_{j=1}^n b_j a_j^{-1}}{\sum_{k=1}^n a_k^{-1}}\right)^2} = \frac{x_i^2\left(x_i^{-1} - \frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)}{\sum_{\ell=1}^n x_\ell^2\left(x_\ell^{-1} - \frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)^2} = \frac{x_i\left(1 - x_i\frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)}{\sum_{\ell=1}^n \left(1 - x_\ell\frac{\sum_{j=1}^n x_j}{\sum_{k=1}^n x_k^2}\right)^2}.$$

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

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

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

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

$\square$

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 $1 \leq i \leq n$. 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 $0 \leq p \leq 1$ 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 7.10, the log-likelihood function has at most one global maximum. So, if the MLE exists, it is unique.

Exercise 7.10. 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.

8. EM Algorithm

Let $X\colon \Omega \to \mathbb{R}^n$ be a discrete or continuous random variable. Let $t\colon \mathbb{R}^n \to \mathbb{R}^m$ be a non-invertible function, and let $Y := t(X)$. For example, let $m < n$, and define t by $t(x_1, \ldots, x_n) := (x_1, \ldots, x_m)\ \forall\ (x_1, \ldots, x_n) \in \mathbb{R}^n$. Suppose we would ideally observe the sample X, but we can only observe the "incomplete" sample Y.

Suppose X has distribution from a family $\{f_\theta\colon \theta \in \Theta\}$ where $f_\theta\colon \mathbb{R}^n \to [0, \infty)$ for all $\theta \in \Theta$. To find the MLE of θ, we would ideally maximize

$$\log \ell(\theta) = \log f_\theta(X).$$

However, since X cannot be directly observed, we cannot compute $\ell(\theta)$ directly, so we might not be able to find the MLE. So, we instead approximate the maximum value of $\log \ell(\theta)$ by conditioning on Y.

The following algorithm tries to find the MLE for Y.

Algorithm 8.1 (**Expectation-Maximization (EM) Algorithm**)**.** Initalize $\theta_0 \in \Theta$. Fix $k \geq 1$. For all $1 \leq j \leq k$, repeat the following procedure:

- (**Expectation**) Given θ_{j-1}, let $\phi_j(\theta) := \mathbf{E}_{\theta_{j-1}}(\log f_\theta(X)|Y)$, for any $\theta \in \Theta$.
- (**Maximization**) Let $\theta_j \in \Theta$ achieve the maximum value of ϕ_j (if it exists).

Remark 8.2. In the case that Y is constant, each step of the algorithm is identical by the Likelihood Inequality, Lemma 2.66. In the case that $Y = X$, the algorithm just outputs the MLE of $Y = X$ in one step. In the case where $m < n$, $X_1, \ldots, X_n\colon \Omega \to \mathbb{R}$ are i.i.d. with common density $f_\theta\colon \mathbb{R} \to [0, \infty)$ and $t(x_1, \ldots, x_n) := (x_1, \ldots, x_m)\ \forall\ (x_1, \ldots, x_n) \in \mathbb{R}^n$, we have

$$\phi_j(\theta) := \mathbf{E}_{\theta_{j-1}}\Big(\sum_{i=1}^{n} \log f_\theta(X_i)\Big|(X_1, \ldots, X_m)\Big) = \sum_{i=1}^{m} \log f_\theta(X_i) + \mathbf{E}_{\theta_{j-1}} \sum_{i=m+1}^{n} \log f_\theta(X_i).$$

So, ϕ_j is the log likelihood for $Y = (X_1, \ldots, X_m)$, plus the expected value of the log likelihood for $X_{m+1}, \ldots, X_n$.

Note that we cannot apply the Likelihood Inequality 2.66 directly to ϕ_j, i.e. the maximum value of ϕ_j is not θ_{j-1}, in general.

Denote $f_{X|Y}(x|y)$ the conditional density (or conditional probability mass function) of X given $Y = y$.

Lemma 8.3. *Suppose X has density f_θ and $Y := t(X)$ has density h_θ. We then denote $g_\theta(x|y) := f_{X|Y}(x|y)$. Then for any $\theta \in \Theta$,*

$$\log h_\theta(Y) - \log h_{\theta_{j-1}}(Y) \geq \phi_j(\theta) - \phi_j(\theta_{j-1}).$$

Equality holds only when $g_\theta(X|y) = g_{\theta_{j-1}}(X|y)$ almost surely with respect to $\mathbf{P}_{\theta_{j-1}}$ (for fixed y).

Proof. Since $f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y)$, we have

$$\log f_Y(y) = \log f_{X,Y}(x, y) - \log f_{X|Y}(x|y).$$

Since $Y = h(X)$, $f_{X,Y}(x, y) = f_X(x) 1_{y=h(x)}$. That is, when $y = h(x)$, we have

$$\log f_Y(y) = \log f_X(x) - \log f_{X|Y}(x|y) = \log f_\theta(x) - \log f_{X|Y}(x|y).$$

Using our streamlined notation, we write instead

$$\log h_\theta(y) = \log f_\theta(x) - \log g_\theta(x|y).$$

Multiplying both sides by $h_{\theta_{j-1}}(x|y)$ and integrating in x, we get

$$\mathbf{E}_{\theta_{j-1}}\Big(\log h_\theta(Y)\Big|Y=y\Big) = \mathbf{E}_{\theta_{j-1}}\Big(\log f_\theta(X)\Big|Y=y\Big) - \mathbf{E}_{\theta_{j-1}}\Big(\log g_\theta(X|y)\Big|Y=y\Big)$$

Setting also $\theta = \theta_{j-1}$ and subtracting one equality from the other, we get

$$\begin{aligned}\log h_\theta(y) - \log h_{\theta_{j-1}}(y) &= \mathbf{E}_{\theta_{j-1}}\Big(\log f_\theta(X)\Big|Y=y\Big) - \mathbf{E}_{\theta_{j-1}}\Big(\log f_{\theta_{j-1}}(X)\Big|Y=y\Big)\\ &\quad - \mathbf{E}_{\theta_{j-1}}\Big(\log g_\theta(X|y)\Big|Y=y\Big) + \mathbf{E}_{\theta_{j-1}}\Big(\log g_{\theta_{j-1}}(X|y)\Big|Y=y\Big)\end{aligned}$$

From the Likelihood Inequality, Lemma 2.66, the sum of the last two terms is nonnegative, and it is zero only when $\log g_\theta(X|y) = \log g_{\theta_{j-1}}(X|y)$ almost surely with respect to $\mathbf{P}_{\theta_{j-1}}$ (for fixed y). In summary,

$$\log h_\theta(Y) - \log h_{\theta_{j-1}}(Y) \geq \phi_j(\theta) - \phi_j(\theta_{j-1}).$$

□

Proposition 8.4 (**EM Algorithm Improvement**). *Let $\theta_0, \ldots, \theta_k$ be an output of the EM Algorithm 8.1. Then for all $1 \leq j \leq k$,*

$$\log h_{\theta_j}(Y) \geq \log h_{\theta_{j-1}}(Y).$$

Proof. By the definition of θ_j in Algorithm 8.1, $\phi_j(\theta_j) \geq \phi_j(\theta_{j-1})$. So, Lemma 8.3 says that

$$\log h_{\theta_j}(Y) - \log h_{\theta_{j-1}}(Y) \geq 0.$$

And equality occurs only when $g_{\theta_j}(X|y) = g_{\theta_{j-1}}(X|y)$ almost surely with respect to $\mathbf{P}_{\theta_{j-1}}$ (for fixed y), or when $\theta_j = \theta_{j-1}$. □

Proposition 8.4 says that the likelihood of Y improves monotonically at each iteration of the EM Algorithm. Moreover, the EM algorithm converges, as the following Theorem demonstrates.

Theorem 8.5 (**EM Algorithm Convergence**). *Fix $y \in \mathbb{R}^m$. Suppose $\Theta \subseteq \mathbb{R}^d$ and $h_\theta(y)$ is a continuous and differentiable function of θ in the interior of Θ. Assume that, for any $\theta_0 \in \Theta$ with $h_{\theta_0}(y) > 0$, $\{\theta \in \Theta\colon h_\theta(y) \geq h_{\theta_0}(y)\}$ is compact and contained in the interior of Θ. Let $\theta_0, \theta_1, \ldots$ be an output of the EM Algorithm 8.1 (that is, we let $k \to \infty$). Then any limit point θ of the sequence $\{\theta_0, \theta_1, \ldots\}$ satisfies $\nabla h_\theta(y) = 0$. (Here ∇ denotes the vector of partial derivatives with respect to $\theta \in \mathbb{R}^d$.) Also, there exists $\theta' \in \Theta$ with $\nabla h_{\theta'}(y) = 0$ such that the sequence $\{h_{\theta_0}(y), h_{\theta_1}(y), \ldots\}$ converges monotonically to $h_{\theta'}(y)$.*

In particular, if h_θ has a unique local maximum, then the EM algorithm converges to this unique (global) maximum.

9. Monte Carlo 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 9.1. 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 9.2 (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 9.3. 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 9.3 then suggest the following method for simulating a random variable on a computer.

Algorithm 9.4 (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 9.5. 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 9.3, F^{-1} is an exponential random variable with parameter 1 if $\mathbf{P}$ is the uniform probability law on $(0,1)$. Or by Algorithm 9.4, $F^{-1}(U)=-\log(1-U)$ is an exponential random variable with parameter 1.

When an explicit formula can be given for Y in Algorithm 9.4, the random variable can be simulated efficiently. However, if Y cannot be accurately or efficiently computed, Algorithm 9.4 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 9.4. 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 1.90, this random variable is close to a standard Gaussian. In fact, explicit error bounds can be given by Theorem 1.98. 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 1.98 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 9.6 (**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

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

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 $e^{(1)},\ldots,e^{(n)}$ be the rows of r. For any $1\leq i\leq n$, define

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

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

$$\mathbf{E}(Z_iZ_j) = a_{ij}, \qquad \forall\, 1\leq i,j\leq n.$$

9.1. **Accept/Reject Sampling.** As mentioned above, one downside of Algorithm 9.4 is that the "inverse" CDF might be difficult to compute directly. For example, the Gaussian inverse CDF has no elementary formula. For continuous random variables, an alternative method of simulation is often easier to use. For simplicity, we state the algorithm first just for continuous random variables with compactly supported PDFs.

Algorithm 9.7 (**Accept/Reject Sampling**). Let $-\infty < a < b < \infty$. Let $f\colon [a,b]\to[0,\infty)$ be the PDF of a real-valued random variable X with maximum value $m := \max_{x\in[a,b]} f(x) < \infty$. Let $n\geq 1$. Let $(X_1,Y_1),(X_2,Y_2),\ldots$ be i.i.d random variables uniformly distributed in the rectangle $[a,b]\times[0,m]$. Define $I := \inf\{n\geq 1\colon Y_n\leq f(X_n)\}$.

Output $Z := X_I$.

In this algorithm, we simulate a uniformly distributed random variable in a rectangle, and we only "accept" the first point (X_i,Y_i) that lies under the graph of the PDF f.

Claim. Z has PDF f.

Proof of Claim. Let $z \in \mathbb{R}$. Define $\varepsilon := \frac{1}{m(b-a)} \int_a^b f(x)dx = \frac{1}{m(b-a)}$. Then

$$\mathbf{P}(Z \leq z) = \sum_{i=1}^{\infty} \mathbf{P}(Z \leq z,\, Z = X_i) = \sum_{i=1}^{\infty} \mathbf{P}(X_i \leq z,\, Y_i \leq f(X_i)) \prod_{j=1}^{i-1} \mathbf{P}(Y_j > f(X_j))$$

$$= \sum_{i=1}^{\infty} \frac{\int_{-\infty}^{z} f(x)dx}{m(b-a)} (1-\varepsilon)^{i-1} = \int_{-\infty}^{z} f(x)dx \sum_{i=1}^{\infty} \varepsilon(1-\varepsilon)^{i-1} = \int_{-\infty}^{z} f(x)dx.$$

□

9.2. **Markov Chain Monte Carlo Introduction.** Algorithm 9.7 works well for continuous random variables with PDFs that have fairly simple formulas. Algorithm 9.4 works well when the CDF or "inverse" CDF have fairly simple formulas. If a discrete random variable has a fairly complicated definition, Algorithms 9.7 and 9.4 may not perform very well. An alternative method to simulate a discrete random variable X of this type involves creating a sequence of random variables that become progressively closer and closer in distribution to X. This sequence of random variables we construct in this way will be a Markov Chain. So, let us now review the theory of Markov Chains.

9.3. **Some Linear Algebra.**

Definition 9.8 (**Eigenvector, Eigenvalue**)**.** Let A be an $m \times m$ real matrix, let $x \in \mathbb{R}^m$ be a column vector, and let $y \in \mathbb{R}^m$ be a row vector. We say x is a (right) **eigenvector** of A with eigenvalue $\lambda \in \mathbb{C}$ if $x \neq 0$ and

$$Ax = \lambda x.$$

We say y is a (left) **eigenvector** of A with eigenvalue $\lambda \in \mathbb{C}$ if $y \neq 0$ and

$$yA = \lambda y.$$

Note that x is a right eigenvector for A if and only if x^T is a left eigenvector of A^T.

Definition 9.9. The **null space** (or **kernel**) of an $m \times n$ real matrix A is the set of all column-vectors $x \in \mathbb{R}^n$ such that $Ax = 0$. The **nullity** of A is the number of nonzero vectors that can form a basis of the null space of A

The **column space** is the set of all linear combinations of the columns of the matrix A. The **rank** of A is the number of nonzero vectors that can form a basis of the column space of A.

Theorem 9.10 (**Rank-Nullity Theorem**)**.** *Let A be an $m \times n$ real matrix. Then the rank of A plus the nullity of A is equal to n.*

9.4. **Markov Chains.** Before defining a Markov chain formally, we give an example of one.

Notation. In the sections below, we denote the sample space as $\mathcal{S}$ and we denote the state space as Ω, since we will often be considering probability distributions on the state space.

Example 9.11 (**Frog on two Lily Pads**)**.** Suppose there are two different lily pads labelled e (for east) and w (for west). Suppose the frog starts on one of the two lily pads. Let $0 < p, q < 1$. There is a coin on the lily pad e which has probability p of landing heads and probability $1 - p$ of landing tails. Similarly, there is a coin on the lily pad w which has probability q of landing heads and probability $1-q$ of landing tails. Every day, the frog flips the coin on the lily pad it currently occupies. If the coin lands heads, the frog goes to the other lily pad. If the coin lands tails, the frog stays on its current lily pad.

For any $n \geq 0$, let X_n be the (random) location of the frog at the beginning of day n. Then the sequence of random variables $X_0, X_1, X_2, \ldots$ describes the sequence of positions that the frog takes. Note that if $\mathcal{S}$ is the sample space, then for any $n \geq 0$, $X_n \colon \mathcal{S} \to \{e, w\}$ is a random variable, taking either the value e or w. We would like to find the probabilities that $X_1, X_2, \ldots$ take the values e and w. To this end, let P be a real 2×2 matrix such that $P(x,y) = \mathbf{P}(X_1 = y \,|\, X_0 = x)$, for all $x, y \in \{e, w\}$. That is,

$$P = \begin{pmatrix} P(e,e) & P(e,w) \\ P(w,e) & P(w,w) \end{pmatrix} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

More generally, note that for any integer $n \geq 1$, $P(x,y) = \mathbf{P}(X_n = y \,|\, X_{n-1} = x)$, since the location of the frog tomorrow only depends on its location today.

Then the random variables $(X_0, X_1, \ldots)$ is a Markov Chain with transition matrix P.

Definition 9.12 (**Finite Markov Chain**). A **finite Markov Chain** is a stochastic process $(X_0, X_1, X_2, \ldots)$ together with a finite set Ω, which is called the **state space** of the Markov Chain, and an $|\Omega| \times |\Omega|$ real matrix P. The random variables $X_0, X_1, \ldots$ take values in the finite set Ω. The matrix P is **stochastic**, that is all of its entries are nonnegative and

$$\sum_{y \in \Omega} P(x,y) = 1, \qquad \forall\, x \in \Omega.$$

And the stochastic process satisfies the following **Markov property**: for all $x, y \in \Omega$, for any $n \geq 1$, and for all events H_{n-1} of the form $H_{n-1} = \cap_{k=0}^{n-1}\{X_k = x_k\}$, where $x_k \in \Omega$ for all $0 \leq k \leq n-1$, such that $\mathbf{P}(H_{n-1} \cap \{X_n = x\}) > 0$, we have

$$\mathbf{P}(X_{n+1} = y \,|\, H_{n-1} \cap \{X_n = x\}) = \mathbf{P}(X_{n+1} = y \,|\, X_n = x) = P(x,y).$$

That is, the next location of the Markov chain only depends on its current location. And the transition probability is defined by $P(x,y)$.

Exercise 9.13. Let P, Q be stochastic matrices of the same size. Show that PQ is a stochastic matrix. Conclude that, if r is a positive integer, then P^r is a stochastic matrix.

Exercise 9.14. Let A, B be events in a sample space. Let $C_1, \ldots, C_n$ be events such that $C_i \cap C_j = \emptyset$ for any $i, j \in \{1, \ldots, n\}$ with $i \neq j$, and such that $\cup_{i=1}^n C_i$ is the whole sample space. Show:

$$\mathbf{P}(A|B) = \sum_{i=1}^n \mathbf{P}(A|B,\, C_i)\mathbf{P}(C_i|B).$$

(Hint: consider using the Total Probability Theorem 1.9 and Proposition 1.57.)

Example 9.15. Returning to the frog example, we have

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

Note that each row of this matrix sums to 1, so P is stochastic. We can then compute the probabilities that X_2 takes various values, by conditioning on the two possible values of X_1.

Using Exercise 9.14, the Markov Property, and the definition of P,

$$\begin{aligned}\mathbf{P}(X_2 = w \,|\, X_0 = e) &= \mathbf{P}(X_2 = w \,|\, X_1 = e, X_0 = e)\mathbf{P}(X_1 = e \,|\, X_0 = e)\\ &\qquad + \mathbf{P}(X_2 = w \,|\, X_1 = w, X_0 = e)\mathbf{P}(X_1 = w \,|\, X_0 = e)\\ &= \mathbf{P}(X_2 = w \,|\, X_1 = e)\mathbf{P}(X_1 = e \,|\, X_0 = e) + \mathbf{P}(X_2 = w \,|\, X_1 = w)\mathbf{P}(X_1 = w \,|\, X_0 = e)\\ &= P(e, w)P(e, e) + P(w, w)P(e, w) = p(1 - p) + (1 - q)p. \qquad (2)\end{aligned}$$

More generally, for any $n \geq 1$, define the 1×2 row vector

$$\mu_n := \big(\mathbf{P}(X_n = e \,|\, X_0 = e), \qquad \mathbf{P}(X_n = w \,|\, X_0 = e)\big).$$

Also, assume the frog starts on the lily pad e, so that $\mu_0 = (1, 0)$. Then (2) generalizes to

$$\mu_n = \mu_{n-1}P, \qquad \forall\, n \geq 1.$$

Iteratively applying this identity,

$$\mu_n = \mu_0 P^n, \qquad \forall\, n \geq 0.$$

What happens when n becomes large? In this case, we might expect the vector μ_n to converge to something as $n \to \infty$. That is, when n becomes very large, the probability that X_n takes a particular value converges to a number. Suppose the vector μ_n converges to some 1×2 row vector π as $n \to \infty$. Note that the entries of μ_n sum to 1 and are nonnegative, so the same is true for π. We claim that

$$\pi = \pi P.$$

That is, π is a (left)-eigenvector of P with eigenvalue 1. To see why $\pi = \pi P$ should be true, note that

$$\pi = \lim_{n\to\infty} \mu_n = \lim_{n\to\infty} \mu_0 P^n = (\lim_{n\to\infty} \mu_0 P^n)P = (\lim_{n\to\infty} \mu_n)P = \pi P.$$

The equation $\pi = \pi P$ allows us to solve for π, since it says

$$\Big(\pi(e), \;\; \pi(w)\Big) = \Big(\pi(e)(1 - p) + \pi(w)q, \;\; \pi(e)p + \pi(w)(1 - q)\Big).$$

So, $0 = -p\pi(e) + \pi(w)q$, $\pi(w) = \pi(e)(p/q)$, and $\pi(e) + \pi(w) = 1$, so $\pi(e)(1 + p/q) = 1$, so

$$\pi(e) = \frac{q}{p + q}, \qquad \pi(w) = \frac{p}{p + q}.$$

That is, when n becomes very large, the frog has probability roughly $q/(q + p)$ of being on the e pad, and it has probability roughly $p/(q + p)$ of being on the w pad.

We can actually say something a bit more precise. For any $n \geq 0$, define

$$\Delta_n = \mu_n(e) - \frac{q}{p + q}.$$

Then, using the definition of μ_{n+1}, and $\mu_n(w) = 1 - \mu_n(e)$, we have, for any $n \geq 0$

$$\Delta_{n+1} = (\mu_n P)(e) - \frac{q}{p + q} = \mu_n(e)(1 - p) + q(1 - \mu_n(e)) - \frac{q}{p + q} = (1 - p - q)\Delta_n.$$

So, iterating this equality, we have

$$\Delta_n = (1 - p - q)^n \Delta_0, \qquad \forall\, n \geq 1.$$

Since $0 < p, q < 1$, this means that the quantity Δ_n is converging exponentially fast to 0. In particular,

$$\lim_{n\to\infty} \Delta_n = 0, \qquad \lim_{n\to\infty} \mu_n = \pi.$$

(A similar argument shows that $\mu_n(w) - \frac{p}{p+q}$ converges exponentially fast to zero)

Exercise 9.16. Let $0 < p, q < 1$. Let $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$. Find the (left) eigenvectors of P, and find the eigenvalues of P. By writing any row vector $x \in \mathbb{R}^2$ as a linear combination of eigenvectors of P (whenever possible), find an expression for xP^n for any $n \geq 1$. What is $\lim_{n\to\infty} xP^n$? Is it related to the vector $\pi = (q/(p+q), p/(p+q))$?

Unfortunately, not all Markov chains converge when n becomes large, as we now demonstrate.

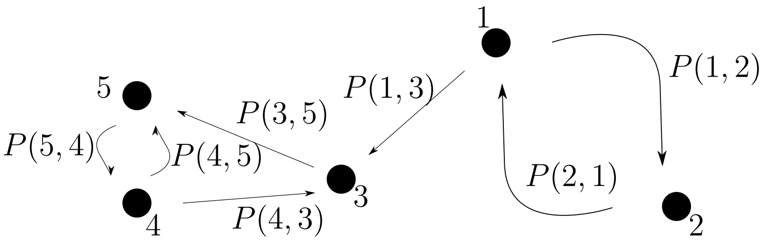
Example 9.17. Consider the Markov chain defined by the matrix $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Note that $P^n = P$ for any positive odd integer n, and $P^n = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ for any positive even integer n. So, if μ is any 1×2 row vector with unequal entries, it is impossible for μP^n to converge as $n \to \infty$.

Example 9.18 (Random Walk on a Graph). A (finite, undirected, simple) **graph** $G = (V, E)$ consists of a finite **vertex set** V and an **edge set** E. The edge set consists of unordered pairs of vertices, so that $E \subseteq \{\{x, y\} \colon x, y \in V, x \neq y\}$. We think of distinct vertices as distinct nodes, where two nodes $x, y \in V$ are joined by an edge if and only if $\{x, y\} \in E$. When $\{x, y\} \in E$, we say that y is a **neighbor** of x (and x is a neighbor of y). The **degree** $\deg(x)$ of a vertex $x \in V$ is the number of neighbors of x. We assume that $\deg(x) > 0$ for every $x \in V$, so that G has no isolated vertices.

Given a graph $G = (V, E)$, we define the **simple random walk** on G to be the Markov chain with state space V and transition matrix

$$P(x, y) = \begin{cases} \frac{1}{\deg(x)} & \text{, if } x \text{ and } y \text{ are neighbors} \\ 0 & \text{, otherwise.} \end{cases}$$

In this Markov chain, starting from any position x, the next state is then any neighbor y of x, each with equal probability. More generally, a **random walk** on a vertex set V is any Markov chain with state space V.



Exercise 9.19. Let $G = (V, E)$ be a graph. Let $|E|$ denote the number of elements in the set E, i.e. $|E|$ is the number of edges of the graph. Prove: $\sum_{x\in V} \deg(x) = 2\,|E|$.

Example 9.20 (**Lazy Random Walk**)**.** Let P be the matrix defined by a simple random walk on a graph $G = (V, E)$. Let I denote the $|V| \times |V|$ identity matrix. The **lazy random walk** is the Markov chain with transition matrix $(P + I)/2$. That is, with probability $1/2$, the next state is your current state, and with probability $1/2$, the next state is any neighbor of the current state, each chosen with equal probability.

Example 9.21 (**Google's PageRank Algorithm**)**.** We can think of the set of all websites on the internet as a graph, where each website is a vertex in V, and $\{x, y\} \in E$ if and only if there is a hyperlink on page x that links to page y (or if there is a hyperlink on page y that links to page x). Let P denote the normalized adjacency matrix, so that $P(x, y) = 1/\deg(x)$ if $\{x, y\} \in E$, and $P(x, y) = 0$ otherwise. Note that P is a stochastic matrix. Let Q be the $|V| \times |V|$ matrix such that all entries of Q are 1. Consider the matrix

$$N := (.85)P + (.15)Q/ |V| .$$

Then N is a stochastic matrix. We can think of the Markov chain associated to N as follows: 85% of the time, you move from one website to another by one of the hyperlinks on that site, each with equal probability. And 15% of the time, you go to any website on the internet, uniformly at random. The PageRank vector π is then a $1 \times |V|$ vector with $\pi(x) \geq 0$ for all $x \in V$, and $\sum_{x\in V} \pi(x) = 1$ such that $\pi = \pi N$. That is, the PageRank value of website $x \in V$ is $\pi(x)$. The most "relevant" websites x have the largest values of $\pi(x)$.

The idea here is that if $\pi(x)$ is large, then the Markov chain will often encounter the website x, so we think of x as being an important website. At the moment, π is not guaranteed to exist. We will return to this issue in Theorem 9.43 below.

9.5. **Classification of States.**

Definition 9.22. Suppose we have a Markov chain $(X_0, X_1, X_2, \ldots)$ with state space Ω. Let $x \in \Omega$ be fixed. For any set A in the sample space, define a probability law $\mathbf{P}_x$ such that

$$\mathbf{P}_x(A) := \mathbf{P}(A|X_0 = x).$$

Similarly, we define $\mathbf{E}_x$ to be the expected value with respect to the probability law $\mathbf{P}_x$.

More generally, if μ is a probability distribution on Ω, we let $\mathbf{P}_\mu$ denote the probability law, given that the Markov chain started from the probability distribution μ, so that $\mathbf{P}(X_0 = x_0) = \mu(x_0)$ for any $x_0 \in \Omega$. So, for example,

$$\mathbf{P}_\mu(X_1 = x_1) = \sum_{x_0\in\Omega} P(x_0, x_1)\mu(x_0), \qquad \forall\, x_1 \in \Omega.$$

Note also that if $x \in \Omega$ is fixed, and if μ is defined so that $\mu(x) = 1$ and $\mu(y) = 1$ for all $y \neq x$, then $\mathbf{P}_\mu = \mathbf{P}_x$.

Definition 9.23 (**Return Time**)**.** Suppose we have a Markov Chain $X_0, X_1, \ldots$ with state space Ω. Let $y \in \Omega$. Define the **first return time** of y to be the following random variable:

$$T_y := \min\{n \geq 1 \colon X_n = y\}.$$

Also, define

$$\rho_{yy} := \mathbf{P}_y(T_y < \infty).$$

That is, ρ_{yy} is the probability that the chain starts at y, and it returns to y in finite time.

Definition 9.24 (**Stopping Time**). A **stopping time** for a Markov chain $X_0, X_1, \ldots$ is a random variable T taking values in $0, 1, 2, \ldots, \cup\{\infty\}$ such that, for any integer $n \geq 0$, the event $\{T = n\}$ is determined by $X_0, \ldots, X_n$. More formally, for any integer $n \geq 1$, there is a set $B_n \subseteq \Omega^{n+1}$ such that $\{T = n\} = \{(X_0, \ldots, X_n) \in B_n\}$. Put another way, the indicator function $1_{\{T=n\}}$ is a function of the random variables $X_0, \ldots, X_n$.

Example 9.25. Fix $y \in \Omega$. The first return time T_y is a stopping time since

$$\begin{aligned}\{T_y = n\} &= \{X_1 \neq y,\, X_2 \neq y, \ldots, X_{n-1} \neq y,\, X_n = y\}\\ &= \{(X_0, \ldots, X_n) \in \Omega \times \{y\}^c \times \cdots \{y\}^c \times \{y\}\}, \qquad \forall\, n \geq 0.\end{aligned}$$

For an intuitive example of a stopping time, suppose $X_0, X_1, \ldots$ is a Markov chain where X_n is the price of a stock at time $n \geq 0$. Then a stopping time could be the first time that the stock price reaches either \$90 or \$100. That is, a stopping time is a stock trading strategy, or a way of "stopping" the random process, but only using information from the past and present. An example of a random variable T that is not a stopping time is to let T be the time that stock price becomes highest, before the price drops to 0. (For example, $\{T = 100\}$ could depend on X_{104}.) So, since T relies on future information, T is not a stopping time.

Theorem 9.26 (**Strong Markov Property**). *Let T be a stopping time for a Markov chain. Let $\ell \geq 1$, and let $A \subseteq \Omega^\ell$. Fix $n \geq 1$. Then, for any $x_0, \ldots, x_n \in \Omega$,*

$$\begin{aligned}\mathbf{P}_{x_0}((X_{T+1}, \ldots, X_{T+\ell}) \in A \,|\, T = n \text{ and } (X_0, \ldots, X_n) = (x_0, \ldots, x_n))&\\ = \mathbf{P}_{x_n}((X_1, \ldots, X_\ell) \in A).&\end{aligned}$$

That is, if we know $T = n$, $X_n = x_n$ and if we know the previous n states of the Markov chain, then this is exactly the same as starting the Markov chain from the state x_n.

Proof. By the definition of the stopping time, there exists $B_n \subseteq \Omega^{n+1}$ such that $\{T = n\} = \{(X_0, \ldots, X_n) \in B_n\}$. If $(x_0, \ldots, x_n) \in B_n$, we then have (using Exercise 9.27)

$$\begin{aligned}&\mathbf{P}_{x_0}((X_{T+1}, \ldots, X_{T+\ell}) \in A \,|\, T = n,\, (X_0, \ldots, X_n) = (x_0, \ldots, x_n))\\ &\quad= \mathbf{P}((X_{T+1}, \ldots, X_{T+\ell}) \in A \,|\, T = n,\, (X_0, \ldots, X_n) = (x_0, \ldots, x_n))\\ &\quad= \mathbf{P}((X_{n+1}, \ldots, X_{n+\ell}) \in A \,|\, T = n,\, (X_0, \ldots, X_n) = (x_0, \ldots, x_n))\\ &\quad= \mathbf{P}((X_{n+1}, \ldots, X_{n+\ell}) \in A \,|\, (X_0, \ldots, X_n) = (x_0, \ldots, x_n))\\ &\quad= \mathbf{P}((X_{n+1}, \ldots, X_{n+\ell}) \in A \,|\, X_n = x_n) \qquad \text{, by Exercise 1.70}\\ &\quad= \mathbf{P}((X_1, \ldots, X_\ell) \in A \,|\, X_0 = x_n) \qquad \text{, by Exercise 1.70}\\ &\quad= \mathbf{P}_{x_n}((X_1, \ldots, X_\ell) \in A), \qquad \text{, by definition of } P_{x_n}.\end{aligned}$$

Finally, if $(x_0, \ldots, x_n) \notin B_n$, then $\{T = n\} \cap \{(X_0, \ldots, X_n) = (x_0, \ldots, x_n)\} = \emptyset$, so the conditional probability of this event is undefined, and there is nothing to prove. □

Exercise 9.27. Let A, B be events such that $B \subseteq \{X_0 = x_0\}$. Then $\mathbf{P}(A|B) = \mathbf{P}_{x_0}(A|B)$.

More generally, if A, B are events, then $\mathbf{P}_{x_0}(A|B) = \mathbf{P}(A|B, X_0 = x_0)$.

Exercise 9.28. Suppose we have a Markov Chain with state space Ω. Let $n \geq 0$, $\ell \geq 1$, let $x_0, \ldots, x_n \in \Omega$ and let $A \subseteq \Omega^\ell$. Using the (usual) Markov property, show that

$$\begin{aligned}\mathbf{P}((X_{n+1}, \ldots, X_{n+\ell}) \in A \,|\, (X_0, \ldots, X_n) = (x_0, \ldots, x_n))&\\ = \mathbf{P}((X_{n+1}, \ldots, X_{n+\ell}) \in A \,|\, X_n = x_n).&\end{aligned}$$

Then, show that

$$\mathbf{P}((X_{n+1},\ldots,X_{n+\ell})\in A\,|\,X_n=x_n)=\mathbf{P}((X_1,\ldots,X_\ell)\in A\,|\,X_0=x_n).$$

(Hint: it may be helpful to use the Multiplication Rule (Proposition 1.8).)

Exercise 9.29. Suppose we have a Markov chain $X_0, X_1, \ldots$ with finite state space Ω. Let $y\in\Omega$. Define $L_y := \max\{n\geq 0\colon X_n=y\}$. Is L_y a stopping time? Prove your assertion.

Example 9.30. If y is in the state space of a Markov chain, recall we defined the return time to be $T_y=\min\{n\geq 1\colon X_n=y\}$. We also verified T_y is a stopping time. Let $T_y^{(1)}=T_y$, and for any $k\geq 2$, define a random variable

$$T_y^{(k)}=\min\{n>T_y^{(k-1)}\colon X_n=y\}.$$

So, $T_y^{(k)}$ is the time of the k^{th} return of the Markov chain to state y. Just as before, we can verify that $T_y^{(k)}$ is a stopping time for any $k\geq 1$.

Let $T:=T_y^{(k-1)}$. Note that if $T<\infty$, then $T_y^{(k)}-T=\min\{n\geq 1\colon X_{T+n}=y\}$. Let $A\subseteq\Omega^\ell$ such that $A=\{y\}^c\times\cdots\times\{y\}^c\times\{y\}$. From the Strong Markov Property (Theorem 9.26), for any $n\geq 1$,

$$\begin{aligned}\mathbf{P}_{x_0}((X_{T+1},\ldots,X_{T+\ell})\in A\,|\,T=n\text{ and }(X_0,\ldots,X_n)=(x_0,\ldots,x_n))\\ =\mathbf{P}_{x_n}((X_1,\ldots,X_\ell)\in A).\end{aligned}$$

Since $\{T_y^{(k)}-T=\ell\}=\{(X_{T+1},\ldots,X_{T+\ell})\in A\}$, and $\{T_y=\ell\}=\{(X_1,\ldots,X_\ell)\in A\}$, if we use $x_0=x_n=y$, we get

$$\mathbf{P}_y(T_y^{(k)}-T=\ell\,|\,T=n,\,X_1=x_1,\ldots,X_{n-1}=x_{n-1},X_n=y)=\mathbf{P}_y(T_y=\ell),\qquad\forall\,\ell,n\geq 1.$$

From the definition of conditional probability,

$$\begin{aligned}&\mathbf{P}_y(T_y^{(k)}-T=\ell,\,T=n,\,X_1=x_1,\ldots,X_{n-1}=x_{n-1},X_n=y)\\ &\quad=\mathbf{P}_y(T=n,\,X_1=x_1,\ldots,X_{n-1}=x_{n-1},X_n=y)\mathbf{P}_y(T_y=\ell)\qquad\forall\,\ell,n\geq 1.\end{aligned}$$

Summing over all $x_1,\ldots,x_{n-1}$ such that $\{X_1=x_1,\ldots,X_{n-1}=x_{n-1},X_n=y\}\subseteq\{T=n\}$,

$$\mathbf{P}_y(T_y^{(k)}-T=\ell,\,T=n)=\mathbf{P}_y(T=n)\mathbf{P}_y(T_y=\ell),\qquad\forall\,\ell,n\geq 1.$$

Taking the union over all $\ell\geq 1$,

$$\mathbf{P}_y(T_y^{(k)}-T<\infty,\,T=n)=\mathbf{P}_y(T=n)\mathbf{P}_y(T_y<\infty)=\mathbf{P}_y(T=n)\rho_{yy},\qquad\forall\,n\geq 1.$$

Then, summing over all $n\geq 1$,

$$\mathbf{P}_y(T_y^{(k)}-T<\infty,\,T<\infty)=\rho_{yy}\mathbf{P}_y(T<\infty).$$

Using the definition of conditional probability again,

$$\mathbf{P}_y(T_y^{(k)}-T<\infty\,|\,T<\infty)=\rho_{yy}.\qquad(*)$$

So, using the multiplication rule (Proposition 1.8) and recalling the definition of T,

$$\begin{aligned}\mathbf{P}_y(T_y^{(k)}<\infty)&=\mathbf{P}_y(T_y^{(k)}-T_y^{(k-1)}<\infty)\\ &=\mathbf{P}_y(T_y^{(k)}-T_y^{(k-1)}<\infty\,|\,T_y^{(k-1)}<\infty)\mathbf{P}_y(T_y^{(k-1)}<\infty)\\ &=\rho_{yy}\mathbf{P}_y(T_y^{(k-1)}<\infty)\qquad,\text{ by }(*)\end{aligned}$$

Iterating this equality $k-1$ times, we have shown:

Proposition 9.31. *For any integer $k \geq 1$,*

$$\mathbf{P}_y(T_y^{(k)} < \infty) = [\mathbf{P}_y(T_y < \infty)]^k = \rho_{yy}^k.$$

In particular, if $\rho_{yy} = 1$, then the Markov chain returns to y an infinite number of times. But if $\rho_{yy} < 1$, then eventually the Markov chain will not return to y:

$$\mathbf{P}_y(T_y^{(k)} = \infty \, \forall \, k \geq j) = \mathbf{P}_y(T_y^{(j)} = \infty) = 1 - \rho_{yy}^j \to 1 \text{ as } j \to \infty.$$

For this reason, we make the following definitions.

Definition 9.32 (Recurrent State, Transient State). If $\rho_{yy} = 1$, we say the state $y \in \Omega$ is **recurrent**. If $\rho_{yy} < 1$, we say the state $y \in \Omega$ is **transient**.

Example 9.33 (Gambler's Ruin). Consider the Markov Chain defined by the following 5×5 stochastic matrix

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ .6 & 0 & .4 & 0 & 0 \\ 0 & .6 & 0 & .4 & 0 \\ 0 & 0 & .6 & 0 & .4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We label the rows and columns of this matrix as $\{1, 2, 3, 4, 5\}$, so that we consider the Markov chain to have state space $\{1, 2, 3, 4, 5\}$. We think of state 1 as a Gambler going bankrupt, state 5 as a Gambler reaching a high amount of money and cashing out. And at each of the states $2, 3, 4$, the gambler can either win a round of some game with probability .4, or lose a round of the game with probability .6.

We will show that states 1 and 5 are recurrent, whereas states $2, 3, 4$ are transient.

Since $P(1, 1) = 1$, $\mathbf{P}_1(T_1 = 1) = 1$, so $\mathbf{P}_1(T_1 < \infty) = 1$. Similarly, $P(5, 5) = 1$, so $\mathbf{P}_5(T_5 = 1)$ and $\mathbf{P}_5(T_5 < \infty) = 1$. So, states 1 and 5 are recurrent.

Now, $P(2, 1) = .6$, and since $P(1, 1) = 1$, if the Markov chain reaches 1 it will never return to 2. So, using the Multiplication rule and the Markov property,

$$\begin{aligned} \mathbf{P}_2(T_2 = \infty) &\geq \mathbf{P}_2(X_1 = 1, X_2 = 1, X_3 = 1, \ldots) \\ &= \mathbf{P}(X_1 = 1 \,|\, X_0 = 2)\mathbf{P}(X_2 = 1 \,|\, X_1 = 1)\mathbf{P}(X_3 = 1 \,|\, X_2 = 1) \cdots \\ &= \lim_{n \to \infty} P(2, 1)P(1, 1)^n = P(2, 1) = .6 > 0. \end{aligned}$$

That is, $\mathbf{P}_2(T_2 < \infty) = 1 - \mathbf{P}(T_2 = \infty) \leq 1 - .6 < 1$, so that state 2 is transient. Similarly, $P(4, 5) = .4$, and $P(5, 5) = 1$, so $\mathbf{P}_4(T_4 = \infty) \geq P(4, 5) > 0$, so $\mathbf{P}_4(T_4 < \infty) < 1$, so state 4 is transient. Using similar reasoning again,

$$\mathbf{P}_3(T_3 = \infty) \geq \lim_{n \to \infty} P(3, 2)P(2, 1)P(1, 1)^n = P(3, 2)P(2, 1) > 0.$$

So, $\mathbf{P}_3(T_3 < \infty) < 1$, so state 3 is transient.

We defined the transition matrix P so that $P(x, y) = \mathbf{P}(X_1 = y \,|\, X_0 = x)$, for any x, y in the state space of the Markov chain. Powers of the matrix P have a similar interpretation. For any $n \geq 1$, $x, y \in \Omega$, define $p^{(n)}(x, y) := \mathbf{P}(X_n = y \,|\, X_0 = x)$.

Proposition 9.34 (**Chapman-Kolmogorov Equation**). *Let $n, m \geq 1$. Let $x, y \in \Omega$ be states of a finite (or countable) Markov chain. Then*

$$p^{(m+n)}(x,y) = \sum_{z\in\Omega} p^{(m)}(x,z)p^{(n)}(z,y)$$

So, for any $x, y, z \in \Omega$, $p^{(m+n)}(x,y) \geq p^{(m)}(x,z)p^{(n)}(z,y)$.

Corollary 9.35. *Let $m \geq 1$. Let $x, y \in \Omega$ be states of a finite Markov chain. Then*

$$P^m(x,y) = p^{(m)}(x,y).$$

Proof of Corollary 9.35. We induct on m. The case $m = 1$ follows since by definition, $p^{(1)}(x,y) = P(x,y)$ for all $x, y \in \Omega$. We now perform the inductive step. From Proposition 9.34 with $n = 1$,

$$p^{(m+1)}(x,y) = \sum_{z\in\Omega} p^{(m)}(x,z)p^{(1)}(z,y) = \sum_{z\in\Omega} P^m(x,z)P(z,y) = P^{m+1}(x,y).$$

The second equality is the inductive hypothesis, and the last equality is the definition of matrix multiplication. □

Proof of Proposition 9.34. Let $x, y \in \Omega$. Using the Total Probability Theorem, we have

$$\begin{aligned}
p^{(m+n)}(x,y) = \mathbf{P}(X_{m+n} = y \,|\, X_0 = x) &= \sum_{z\in\Omega} \mathbf{P}(X_{m+n} = y, X_m = z \,|\, X_0 = x) \\
&= \sum_{z\in\Omega} \frac{\mathbf{P}(X_{m+n} = y, X_m = z, X_0 = x)}{\mathbf{P}(X_0 = x)} \\
&= \sum_{z\in\Omega} \frac{\mathbf{P}(X_{m+n} = y, X_m = z, X_0 = x)}{\mathbf{P}(X_m = z, X_0 = x)} \frac{\mathbf{P}(X_m = z, X_0 = x)}{\mathbf{P}(X_0 = x)} \\
&= \sum_{z\in\Omega} \mathbf{P}(X_{m+n} = y \,|\, X_m = z, X_0 = x)\mathbf{P}(X_m = z \,|\, X_0 = x).
\end{aligned}$$

Finally, the Markov property and Exercise 9.28 imply that

$$\begin{aligned}
p^{(m+n)}(x,y) &= \sum_{z\in\Omega} \mathbf{P}(X_{m+n} = y \,|\, X_m = z)\mathbf{P}(X_m = z \,|\, X_0 = x) \\
&= \sum_{z\in\Omega} \mathbf{P}(X_n = y \,|\, X_0 = z)\mathbf{P}(X_m = z \,|\, X_0 = x) = \sum_{z\in\Omega} p^{(n)}(z,y)p^{(m)}(x,z).
\end{aligned}$$

(Since we only condition on events with positive probability, we did not divide by zero.) □

Definition 9.36 (**Irreducible**). A Markov chain with state space Ω and with transition matrix P is called **irreducible** if, for any $x, y \in \Omega$, there exists an integer $n \geq 1$ (which is allowed to depend on x, y) such that $P^n(x,y) > 0$. That is the Markov chain is irreducible if any state can reach any other state, with some positive probability, if the chain runs long enough.

Lemma 9.37. *Suppose we have a finite irreducible Markov chain with state space Ω. Then there exists $0 < \alpha < 1$ and there exists an integer $j > 0$ such that, for any $x, y \in \Omega$,*

$$\mathbf{P}_x(T_y > kj) \leq \alpha^k, \qquad \forall\, k \geq 1.$$

Proof. As a consequence of irreducibility, there exists $\varepsilon > 0$ and integer $j > 0$ such that, for any $x, y \in \Omega$, there exists $r(x,y) \leq j$ such that $P^{r(x,y)}(x,y) > \varepsilon$. That is, after at most j steps of the Markov chain, the chain will move from x to y with some positive probability.

$$\begin{aligned}
\mathbf{P}_x(T_y > kj) &= \mathbf{P}_x(T_y > kj \,|\, T_y > (k-1)j)\mathbf{P}_x(T_y > (k-1)j) \\
&\leq \max_{z\in\Omega} \mathbf{P}_z(T_y > j)\mathbf{P}_x(T_y > (k-1)j), \qquad \text{by Exercise 9.38} \\
&\leq \max_{z\in\Omega} \mathbf{P}_z(T_y > r(z,y))\mathbf{P}_x(T_y > (k-1)j), \qquad \text{since } r(z,y) \leq j \\
&= \max_{z\in\Omega} (1 - \mathbf{P}_z(T_y \leq r(z,y)))\mathbf{P}_x(T_y > (k-1)j) \\
&\leq \max_{z\in\Omega} (1 - P^{r(z,y)}(z,y))\mathbf{P}(T_y > (k-1)j), \qquad \text{by Exercise 9.39} \\
&\leq (1-\varepsilon)\mathbf{P}(T_y > (k-1)j).
\end{aligned}$$

Iterating this inequality $k-1$ times concludes the Lemma with $\alpha := 1 - \varepsilon$. □

Exercise 9.38. Let x, y be points in the state space of a finite Markov Chain $(X_0, X_1, \ldots)$. Let $T_y = \min\{n \geq 1 \colon X_n = y\}$ be the first arrival time of y. Let j, k be positive integers. Show that

$$\mathbf{P}_x(T_y > kj \,|\, T_y > (k-1)j) \leq \max_{z\in\Omega} \mathbf{P}_z(T_y > j).$$

(Hint: use Exercise 9.28)

Exercise 9.39. Let x, y be points in the state space of a finite Markov Chain $(X_0, X_1, \ldots)$ with transition matrix P. Let $T_y = \min\{n \geq 1 \colon X_n = y\}$ be the first arrival time of y. Let j be a positive integer. Show that

$$P^j(x,y) \leq \mathbf{P}_x(T_y \leq j).$$

(Hint: can you induct on j?)

Example 9.40. Consider the Markov Chain with state space $\Omega = \{1, 2, 3\}$ and transition matrix

$$P = \begin{pmatrix} .2 & .3 & .5 \\ .3 & .3 & .4 \\ .4 & .5 & .1 \end{pmatrix}.$$

Then for any x, y in the state space of the Markov chain, $P(x,y) \geq .1$. So, we can use $j = r = 1$ and $\varepsilon = .1$, $\alpha = .9$ in Lemma 9.37 to get

$$\mathbf{P}_x(T_y > k) \leq (.9)^k, \qquad \forall\, k \geq 1, \forall\, x, y \in \Omega.$$

In particular, $\mathbf{P}_y(T_y < \infty) = 1$, so all states are recurrent.

Exercise 9.41. Let x, y be any states in a finite irreducible Markov chain. Show that $\mathbf{E}_x T_y < \infty$. In particular, $\mathbf{P}_y(T_y < \infty) = 1$, so all states are recurrent.

9.6. **Stationary Distribution.**

Definition 9.42 (**Stationary Distribution**). Let P be the $m \times m$ transition matrix of a finite Markov chain with state space Ω. Let π be a $1 \times m$ row vector. We say that π is a **stationary distribution** if $\pi(x) \geq 0$ for every $x \in \Omega$, $\sum_{x\in\Omega} \pi(x) = 1$, and if π satisfies

$$\pi = \pi P.$$

As discussed above, if a stationary distribution exists, we can think of $\pi(x)$ as roughly the fraction of time that the Markov chain spends in x, when the Markov chain runs for a long period of time. Put another way, after the Markov chain has run for a long period of time, $\pi(x)$ is the probability that the Markov chain is in state x. In fact, π defines a probability law on the state space Ω: for any $A \subseteq \Omega$, define $\pi(A) := \sum_{x\in A}\pi(x)$. Then π is a probability law on Ω.

Unfortunately, even if the stationary distribution exists, it may not be unique! If there is more than one stationary distribution, then there may not be a sensible way of describing where the Markov chain could be, after a long time has passed.

In this section, we address the existence and uniqueness of a stationary distribution π.

Theorem 9.43 (**Existence**)**.** *Suppose we have a finite irreducible Markov chain $(X_0, X_1, \ldots)$ with state space Ω and transition matrix P. Then there exists a stationary distribution π such that $\pi = \pi P$ and $\pi(x) > 0$ for all $x \in \Omega$.*

Proof. Let $y, z \in \Omega$. Let let $T_z = \min\{n \geq 1 \colon X_n = z\}$. We define $\widetilde{\pi}(y)$ to be the expected number of times the chain visits y before returning to z. That is, define

$$\widetilde{\pi}(y) = \mathbf{E}_z\left(\sum_{n=0}^{\infty} 1_{\{X_n=y,\, T_z>n\}}\right) = \sum_{n=0}^{\infty}\mathbf{P}_z(X_n = y,\, T_z > n). \qquad (*)$$

First, note that since the Markov chain is irreducible, there is always some probability that the chain starts at z and visits y before returning to z. Therefore, $\widetilde{\pi}(y) > 0$ for any $y \in \Omega$. Now, using Remark 1.34, and then Exercise 9.41,

$$\widetilde{\pi}(y) \leq \sum_{n=0}^{\infty}\mathbf{P}_z(T_z > n) = \mathbf{E}_z T_z < \infty, \qquad \forall\, y \in \Omega.$$

We now show that $\widetilde{\pi}$ satisfies $\widetilde{\pi} = \widetilde{\pi}P$. By definition of $\widetilde{\pi}$,

$$\sum_{x\in\Omega}\widetilde{\pi}(x)P(x,y) = \sum_{x\in\Omega}\sum_{n=0}^{\infty}\mathbf{P}_z(X_n = x,\, T_z > n)P(x,y). \qquad (**)$$

Consider the event $\{T_z > n\} = \{T_z \geq n+1\} = \{T_z \leq n\}^c$. That is, $\{T_z > n\}$ only depends on $X_0, \ldots, X_n$. So, the usual Markov property (rearranged a bit) says

$$\mathbf{P}_z(X_{n+1} = y,\, X_n = x,\, T_z \geq n+1) = \mathbf{P}_z(X_n = x,\, T_z \geq n+1)P(x,y).$$

Substituting this into $(**)$ and first changing the order of summation,

$$\begin{aligned}\sum_{x\in\Omega}\widetilde{\pi}(x)P(x,y) &= \sum_{n=0}^{\infty}\sum_{x\in\Omega}\mathbf{P}_z(X_{n+1} = y,\, X_n = x,\, T_z \geq n+1)\\ &= \sum_{n=0}^{\infty}\mathbf{P}_z(X_{n+1} = y,\, T_z \geq n+1) = \sum_{n=1}^{\infty}\mathbf{P}_z(X_n = y,\, T_z \geq n)\\ &= \widetilde{\pi}(y) - \mathbf{P}_z(X_0 = y,\, T_z > 0) + \sum_{n=1}^{\infty}\mathbf{P}_z(X_n = y,\, T_z = n), \qquad \text{by } (*)\\ &= \widetilde{\pi}(y) - \mathbf{P}_z(X_0 = y) + \mathbf{P}_z(X_{T_z} = y), \qquad \text{substituting } n = T_z.\end{aligned}$$

We now split into two cases. If $y = z$, then $\mathbf{P}_z(X_0 = y) = 1$ by definition of $\mathbf{P}_z$, and also $X_{T_z} = z = y$ by definition of T_z, so $\mathbf{P}_z(X_{T_z} = y) = 1$. If $y \neq z$, then by similar reasoning, $\mathbf{P}_z(X_0 = y) = \mathbf{P}_z(X_{T_z} = y) = 0$ In any case $-\mathbf{P}_z(X_0 = y,\, T_z > 0) + \mathbf{P}_z(X_{T_z} = y) = 0$. In conclusion, we have shown that

$$\widetilde{\pi} = \widetilde{\pi} P.$$

Finally, to get a stationary distribution π also satisfying $\pi = \pi P$, we just define $\pi(x) := \widetilde{\pi}(x)/\sum_{y\in\Omega}\widetilde{\pi}(y)$ for any $x \in \Omega$. □

Remark 9.44. We note in passing the following identity. By $(*)$ and Remark 1.34,

$$\sum_{y\in\Omega}\widetilde{\pi}(y) = \sum_{n=0}^{\infty}\sum_{y\in\Omega}\mathbf{P}_z(X_n = y,\, T_z > n) = \sum_{n=0}^{\infty}\mathbf{P}_z(T_z > n) = \mathbf{E}_z T_z.$$

Lemma 9.45. *Let P be the transition matrix of a finite irreducible Markov chain with state space Ω. Let $f\colon \Omega \to \mathbb{R}$ be a **harmonic** function, so that*

$$f(x) = \sum_{y\in\Omega} P(x,y)f(y), \qquad \forall\, x \in \Omega.$$

Then f is a constant function.

Proof. Since Ω is finite, there exists $x_0 \in \Omega$ such that $M := \max_{x\in\Omega} f(x) = f(x_0)$. Let $z \in \Omega$ with $P(x_0, z) > 0$, and assume that $f(z) < M$. Then since f is harmonic,

$$f(x_0) = P(x_0,z)f(z) + \sum_{y\in\Omega\colon\, y\neq z} P(x_0,y)f(y) < M\sum_{y\in\Omega}P(x_0,y) = M,$$

a contradiction. Thus, $f(z) = M$ for any $z \in \Omega$ with $P(x_0, z) > 0$.

Finally, for any $z \in \Omega$, irreducibility of P implies that there is a sequence of points $x_0, x_1, \ldots, x_k = z$ in Ω such that $P(x_i, x_{i+1}) > 0$ for every $0 \leq i < k$. So, by repeating the above argument $k - 1$ times, $M = f(x_0) = f(x_1) = \cdots = f(x_k) = f(z)$. That is, $f(z) = M$ for every $z \in \Omega$. □

Theorem 9.46 (Uniqueness). *Let P be the transition matrix of a finite irreducible Markov chain. Then there exists a unique stationary distribution π such that $\pi = \pi P$.*

Proof. By Theorem 9.43, there exists at least one stationary distribution π such that $\pi = \pi P$. Let I denote the $|\Omega| \times |\Omega|$ identity matrix. Lemma 9.45 implies that the null-space of $P - I$ has dimension 1. So, by the rank-nullity theorem, the column rank of $P - I$ is $|\Omega| - 1$. Since row rank and column rank are equal, the row rank of $P - I$ is $|\Omega| - 1$. That is, the space of solutions of the row-vector equation $\mu = \mu P$ is one-dimensional (where μ denotes a $1 \times |\Omega|$ row vector.) Since this space is one-dimensional, it has only one vector whose entries sum to 1. □

The following Corollary gives a sensible way of computing the stationary distribution of an irreducible Markov chain.

Corollary 9.47. *Let P be the transition matrix of a finite irreducible Markov chain with state space Ω. If π is the unique solution to $\pi = \pi P$, then*

$$\pi(x) = \frac{1}{\mathbf{E}_x T_x}, \qquad \forall\, x \in \Omega.$$

Proof. Let $y, z \in \Omega$ and define $\widetilde{\pi}_z(y) := \widetilde{\pi}(y)$, where $\widetilde{\pi}(y)$ is defined in $(*)$ in Theorem 9.43. Also, define $\pi_z(y) := \widetilde{\pi}_z(y)/\mathbf{E}_z T_z$. Theorem 9.43 and Remark 9.44 imply that π_z is a stationary distribution such that $\pi_z = \pi_z P$. Theorem 9.46 implies that π_z does not depend on z. That is, for any $x \in \Omega$, if we define $\pi(x) := \pi_z(x)$ (for any particular $z \in \Omega$, since the expression does not depend on z), then we have $\pi = \pi P$, and

$$\pi(x) = \pi_x(x) = \frac{\widetilde{\pi}_x(x)}{\mathbf{E}_x T_x} = \frac{1}{\mathbf{E}_x T_x}.$$

In the last equality, we used $\widetilde{\pi}_x(x) = 1$, which follows by the definition of $\widetilde{\pi}_x$. (The $n = 0$ term in $\sum_{n=0}^{\infty} \mathbf{P}_x(X_n = x,\, T_x > n)$ is 1, and all other terms in the sum are zero.) □

Exercise 9.48 (Knight Moves)**.** Consider a standard 8×8 chess board. Let V be a set of vertices corresponding to each square on the board (so V has 64 elements). Any two vertices $x, y \in V$ are connected by an edge if and only if a knight can move from x to y. (The knight chess piece moves in an L-shape, so that a single move constitutes two spaces moved along the horizontal axis followed by one move along the vertical axis (or two spaces moved along the vertical axis, followed by one move along the horizontal axis.) Consider the simple random walk on this graph. This Markov chain then represents a knight randomly moving around a chess board. For every space x on the chessboard, compute the expected return time $\mathbf{E}_x T_x$ for that space. (It might be convenient to just draw the expected values on the chessboard itself.)

Exercise 9.49 (Simplified Monopoly)**.** Let $\Omega = \{1, 2, \ldots, 10\}$. We consider Ω to be the ten spaces of a circular game board. You move from one space to the next by rolling a fair six-sided die. So, for example $P(1, k) = 1/6$ for every $2 \leq k \leq 7$. More generally, for every $j \in \Omega$ with $j \neq 5$, $P(j, k) = 1/6$ if $k = (j + i) \bmod 10$ for some $1 \leq i \leq 6$. Finally, the space 5 forces you to return to 1, so that $P(5, 1) = 1$. (Note that $\bmod 10$ denotes arithmetic modulo 10, so e.g. $7 + 5 = 2 \bmod 10$.)

Using a computer, find the unique stationary distribution of this Markov chain. Which point has the highest stationary probability? The lowest?

Compare this stationary distribution to the stationary distribution that arises from the doubly stochastic matrix: for all $j \in \Omega$, $P(j, k) = 1/6$ if $k = (j+i) \bmod 10$ for some $1 \leq i \leq 6$. (See Exercise 9.52.)

Exercise 9.50. Give an example of a Markov chain where there are at least two different stationary distributions.

Exercise 9.51. Is there a finite Markov chain where no stationary distribution exists? Either find one, or prove that no such finite Markov chain exists.

(If you want to show that no such finite Markov chain exists, you are allowed to just prove the weaker assertion that: for every stochastic matrix P, there always exists a nonzero vector π with $\pi = \pi P$.)

Exercise 9.52. Let P be the transition matrix for a finite Markov chain with state space Ω. We say that the matrix P is **doubly stochastic** if the columns of P each sum to 1. (Since P is a transition matrix, each of its rows already sum to 1.) Let π such that $\pi(x) = 1/|\Omega|$ for all $x \in \Omega$. That is, π is uniform on Ω. Show that $\pi = \pi P$.

Remark 9.53. If a finite Markov chain is not irreducible, we can divide the state space into pieces, each of which is irreducible (or transient), and then study how the Markov chain acts on each individual piece. (For a precise statement, see Theorem 1.8 in the Durrett book.)

Definition 9.54 (Reversible). Let P be the transition matrix of a finite Markov chain with state space Ω. We say that the Markov chain is **reversible** if there exists a probability distribution π on Ω satisfying the following **detailed balance condition**:

$$\pi(x)P(x,y) = \pi(y)P(y,x), \qquad \forall\, x, y \in \Omega.$$

Exercise 9.55. Give an example of a random walk on a graph that is not reversible.

Proposition 9.56 (Reversible Implies Stationary). *Let π be a probability distribution satisfying the detailed balance condition for a finite Markov chain. Then π is a stationary distribution.*

Proof. We sum both sides of the detailed balance condition over y, and use that P is stochastic to get

$$(\pi P)(x) = \sum_{y\in\Omega} \pi(y)P(y,x) = \pi(x)\sum_{y\in\Omega} P(x,y) = \pi(x).$$

□

Exercise 9.57. Let P be the transition matrix of a finite, irreducible, reversible Markov chain with state space Ω and stationary distribution π. Let $f, g \in \mathbb{R}^{|\Omega|}$ be column vectors. Consider the following bilinear function on f, g, which is referred to as an inner product (or dot product):

$$\langle f, g\rangle := \sum_{x\in\Omega} f(x)g(x)\pi(x).$$

Show that P is self-adjoint (i.e. symmetric) in the sense that

$$\langle f, Pg\rangle = \langle Pf, g\rangle.$$

In particular (for those that have taken 115A), the spectral theorem implies that all eigenvalues of P are real.

Finally, find a transition matrix P such that at least one eigenvalue of P is not real.

Proposition 9.58. *Suppose we have a finite irreducible Markov chain with state space Ω, transition matrix P and stationary distribution π. Fix $n \geq 1$, and for any $0 \leq m \leq n$, define $\widehat{X}_m = X_{n-m}$. Then $\widehat{X}_m$ is a Markov chain with transition probabilities given by*

$$\widehat{P}(x,y) = \frac{\pi(y)P(y,x)}{\pi(x)}, \qquad \forall\, x, y \in \Omega.$$

Moreover, π is stationary for $\widehat{P}$, and we have

$$\mathbf{P}_\pi(X_0 = x_0, \ldots, X_n = x_n) = \mathbf{P}_\pi(\widehat{X}_0 = x_n, \ldots, \widehat{X}_n = x_0), \qquad \forall\, x_0, \ldots, x_n \in \Omega.$$

Proof. First, from Theorem 9.43, $\pi(x) > 0$ for all x in the state space of the Markov chain, so we have not divided by zero. Now, we first check π is stationary for $\widehat{P}$:

$$\sum_{y\in\Omega} \pi(y)\widehat{P}(y,x) = \sum_{y\in\Omega} \pi(y)\frac{\pi(x)P(x,y)}{\pi(y)} = \pi(x).$$

Using similar reasoning, we know that $\sum_{y\in\Omega}\widehat{P}(x,y)=1$, so that $\widehat{P}$ is itself a stochastic matrix. Finally, noting that $P(x_{i-1},x_i)=\pi(x_i)\widehat{P}(x_i,x_{i-1})/\pi(x_{i-1})$ for each $1\leq i\leq n$,

$$\begin{aligned}\mathbf{P}_\pi(X_0=x_0,\ldots,X_n=x_n)&=\pi(x_0)P(x_0,x_1)\cdots P(x_{n-1},x_n)\\&=\pi(x_n)\widehat{P}(x_n,x_{n-1})\cdots\widehat{P}(x_1,x_0)\\&=\mathbf{P}_\pi(\widehat{X}_0=x_n,\ldots,\widehat{X}_n=x_0).\end{aligned}$$

□

Remark 9.59. If the Markov chain is reversible, then $\widehat{P}=P$. So, being reversible means that the Markov chain can be run backwards or forwards in the same way, if we start the Markov chain from the stationary distribution.

Example 9.60. We return to Example 9.18. Let $G=(V,E)$ be a graph with at least one edge, and let P correspond to the simple random walk on G. So, $P(x,y)=1/\deg(x)$ if x and y are neighbors, and $P(x,y)=0$ otherwise. For any $x\in V$, define $\pi(x):=\deg(x)/(2\,|E|)$. We show π is stationary. From Proposition 9.56, it suffices to show the detailed balance condition holds.

If x and y are not neighbors, then $P(x,y)=P(y,x)=0$, and both sides of the detailed balance condition are equal. If x and y are neighbors, then

$$\pi(x)P(x,y)=\frac{\deg(x)}{2\,|E|}\frac{1}{\deg(x)}=\frac{1}{2\,|E|}=\frac{\deg(y)}{2\,|E|}\frac{1}{\deg(y)}=\pi(y)P(y,x).$$

Exercise 9.61 (Ehrenfest Urn Model). Suppose we have two urns and n spheres. Each sphere is in either of the first or the second urn. At each step of the Markov chain, one of the spheres is chosen uniformly at random and moved from its current urn to the other urn. Let X_n be the number of spheres in the first urn at time n. A state of the Markov chain is an integer in $\{0,1,\ldots,n\}$, which represents the number of spheres in the first urn. Then for any $j,k\in\{1,\ldots,n\}$, the transition matrix defining the Markov chain is

$$P(j,k)=\begin{cases}\frac{n-j}{n} & \text{, if } k=j+1\\ \frac{j}{n} & \text{, if } k=j-1\\ 0 & \text{, otherwise.}\end{cases}$$

Show that the unique stationary distribution for this Markov chain is a binomial PMF with parameters n and $1/2$.

Exercise 9.62. Let $V=\{0,1\}^n$ be a set of vertices. We construct a graph from V as follows. Let $x=(x_1,\ldots,x_n),y=(y_1,\ldots,y_n)\in\{0,1\}^n$. Then x and y are connected by an edge in the graph if and only if $\sum_{i=1}^n|x_i-y_i|=1$. That is, x and y are connected if and only if they differ by a single coordinate.

For any $x\in V$, define $f(x)=\sum_{i=1}^n x_i$, $f\colon V\to\{0,1,\ldots,n\}$. Given $x\in V$, we identify x with the state in the Ehrenfest urn model where the first urn has exactly $f(x)$ spheres. Show that the Ehrenfest urn model is a **projection** of the simple random walk on V in the following sense. The probability that $x\in V$ transitions to any state $z\in V$ such that $y=f(z)$ is equal to: the probability that Ehrenfest model with state $f(x)$ transitions to state y.

Moreover, the unique stationary distribution for the simple random walk on V can be projected to give the unique stationary distribution in the Ehrenfest model. That is, if π is the unique stationary distribution for the simple random walk on V, and if for any $A \subseteq \{0, 1, \ldots, n\}$, we define $\mu(A) := \pi(f^{-1}(A))$, then μ is a Binomial PMF with parameters n and $1/2$. (Here $f^{-1}(A) = \{x \in V : f(x) \in A\}$.)

Exercise 9.63 (**Birth-and-Death Chains**)**.** A birth-and-death chain can model the size of some population of organisms. Fix a positive integer k. Consider the state space $\Omega = \{0, 1, 2, \ldots, k\}$. The current state is the current size of the population, and at each step the size can increase or decrease by at most 1. We define $\{(p_n, r_n, q_n)\}_{n=0}^{k}$ such that $p_n + r_n + q_n = 1$ and $p_n, r_n, q_n \geq 0$ for each $0 \leq n \leq k$, and

- $P(n, n+1) = p_n > 0$ for every $0 \leq n < k$.
- $P(n, n-1) = q_n > 0$ for every $0 < n \leq k$.
- $P(n, n) = r_n \geq 0$ for every $0 \leq n \leq k$.
- $q_0 = p_k = 0$.

Show that the birth-and-death chain is reversible.

9.7. **Limiting Behavior.** From Theorem 9.46, we know an irreducible Markov chain has a unique stationary distribution, and Corollary 9.47 gives a sensible way of computing that stationary distribution. But what does this distribution tell us about the Markov chain's behavior? In general, it might not say anything! For example, recall Example 9.17, where we considered the transition matrix $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. If $\mu = (\mu(1), \mu(2))$ is any 1×2 row vector, then $\mu P^n = \mu$ for n even, and $\mu P^n = (\mu(2), \mu(1))$ for n odd. So, if the Markov chain starts at the probability distribution μ where $\mu(1) \neq \mu(2)$, then it is impossible for $\lim_{n\to\infty} \mu P^n$ to exist. That is, there is no sensible way of talking about the limiting behavior of this Markov chain.

Put another way, we need to eliminate this "periodic" behavior to hope to get convergence of the Markov chain. Thankfully, if an irreducible Markov chain has no "periodic" behavior as in the above example, then it does actually converge as $n \to \infty$. In fact, we will be able to give an exponential rate of convergence of the Markov chain. Before doing so, we formally define periodic behavior, and we formally define periodicity and how the Markov chain converges.

Definition 9.64 (**Period, Aperiodic**)**.** Let P be the transition matrix of a finite Markov chain with state space Ω. For any $x \in \Omega$, let $\mathcal{N}(x) := \{n \geq 1 : P^n(x, x) > 0\}$. The **period** of state $x \in \Omega$ is the largest integer that divides all of the integers in $\mathcal{N}(x)$. That is, the period of x, denoted $\gcd \mathcal{N}(x)$, is the greatest common divisor of $\mathcal{N}(x)$. (If $\mathcal{N}(x) = \emptyset$, we leave $\gcd \mathcal{N}(x)$ undefined.) (We say an integer m divides an integer n if there exists an integer k such that $n = km$.)

A Markov chain is called **aperiodic** if all $x \in \Omega$ have period 1.

Exercise 9.65. Give an explicit example of a Markov chain where every state has period 100.

Lemma 9.66. *Let P be the transition matrix of an irreducible, finite Markov chain with state space Ω. Then $\gcd \mathcal{N}(x) = \gcd \mathcal{N}(y)$ for all $x, y \in \Omega$.*

Proof. Let $x, y \in \Omega$. Since the Markov chain is irreducible, there exist $r, \ell \geq 1$ such that $P^r(x,y) > 0$ and $P^\ell(y,x) > 0$. Let $m = r + \ell$. Then $m \in \mathcal{N}(x) \cap \mathcal{N}(y)$ (since $P^m(x,x) \geq P^r(x,y)P^\ell(y,x) > 0$, and $P^m(y,y) \geq P^\ell(y,x)P^r(x,y) > 0$), and $\mathcal{N}(x) \subseteq \mathcal{N}(y) - m$. (If $P^k(x,x) > 0$, then $P^{k+m}(y,y) \geq P^\ell(y,x)P^k(x,x)P^r(x,y) > 0$.) Since $\gcd \mathcal{N}(y)$ divides m and all elements of $\mathcal{N}(y)$, we conclude that $\gcd \mathcal{N}(y)$ divides all elements of $\mathcal{N}(x)$. In particular, $\gcd \mathcal{N}(y) \leq \gcd \mathcal{N}(x)$. Reversing the roles of x and y in the above argument, $\gcd \mathcal{N}(x) \leq \gcd \mathcal{N}(y)$. □

Lemma 9.67. *Let P be the transition matrix of an aperiodic, irreducible, finite Markov chain with state space Ω. Then there exists an integer $r > 0$ such that $P^r(x,y) > 0$ for all $x, y \in \Omega$. (That is, we can choose the r to not depend on x, y.)*

Proof. Since the Markov chain is aperiodic, $\gcd \mathcal{N}(x) = 1$. The set $\mathcal{N}(x)$ is closed under addition, since if $n, m \in \mathcal{N}(x)$, then $P^{n+m}(x,x) \geq P^n(x,x)P^m(x,x) > 0$, so that $n + m \in \mathcal{N}(x)$. From Lemma 9.68 with $g = 1$, there exists $n(x)$ such that if $n \geq n(x)$, then $n \in \mathcal{N}(x)$. Since the Markov chain is irreducible, for any $y \in \Omega$ there exists $r = r(x,y)$ such that $P^r(x,y) > 0$. So, if $n \geq n(x) + r$, we have

$$P^n(x,y) \geq P^{n-r}(x,x)P^r(x,y) > 0.$$

So, if $n \geq n'(x) := n(x) + \max_{x,y\in\Omega} r(x,y)$, then $P^n(x,y) > 0$ for all $y \in \Omega$. Then, if $n \geq \max_{x\in\Omega} n'(x)$, then $P^n(x,y) > 0$ for all $x, y \in \Omega$. □

Lemma 9.68. *Let S be a nonempty subset of the positive integers. Let $g = \gcd(S)$. Then there exists some integer n_S such that, for all $m \geq n_S$, the product mg can be written as a linear combination of elements of S, with nonnegative integer coefficients.*

Proof. Let g^* be the smallest positive integer which is an integer combination of elements of S. Then $g^* \leq s$ for every $s \in S$. Also, g^* divides every element of S (if $s \in S$ and if g^* does not divide s, then the remainder obtained by dividing s by g^* would be smaller than g^*, while being an integer combination of elements of S). So, $g^* \leq g$. Since g divides every element of S as well, g divides g^*, and $g \leq g^*$. So, $g = g^*$.

Now, without loss of generality, we can assume S is finite, since the case that S is infinite follows from the case that S is finite. The case when S has one element is clear. As a base case, we consider when $S = \{a, b\}$, where a, b are distinct positive integers. Let $m > 0$. Since $g = g^*$ and $mg \geq g^*$, we can write $mg = ca + db$ for some integers c, d. Since $mg = ca + db$, we can also write $mg = (c+kb)a + (d-ka)b$ for any k. That is, we can write $mg = ca + db$ for integers c, d with $0 \leq c \leq b-1$. If $mg > (b-1)a - b$, then $db = mg - ca \geq mg - a(b-1) > -b$. So, $d \geq 0$ as well. That is, we can choose n_S such that $n_S \geq ((ab - a - b)/g) + 1$.

We now induct on the size of S, by adding one element a to S. Let $g_S := \gcd(S)$ and let $g := \gcd(\{a\} \cup S)$. For any positive integer a, the definition of gcd implies that $\gcd(\{a\} \cup S) = \gcd(a, g_S)$. Suppose m satisfies $mg \geq n_{\{a,g_S\}}g + n_S g_S$. Then we can write $mg - n_S g_S = ca + dg_S$ for integers $c, d \geq 0$, from the case when S could be $\{a, g_S\}$. Therefore, $mg = ca + (d + n_S)g_S = ca + \sum_{s\in S} c_s s$ for some integers $c_s \geq 0$, by definition of n_S, and using $d + n_S \geq n_S$. In conclusion, we can choose $n_{\{a\}\cup S} = n_{\{a,g_S\}} + n_S g_S/g$, completing the inductive step. □

Definition 9.69 (**Total Variation Distance**)**.** Let μ, ν be probability distributions on a finite state space Ω. We define the **total variation distance** between μ and ν to be

$$\|\mu - \nu\|_{\text{TV}} := \max_{A \subseteq \Omega} |\mu(A) - \nu(A)|.$$

Exercise 9.70. Let Ω be a finite state space. This exercise demonstrates that the total variation distance is a metric. That is, the following three properties are satisfied:

- $\|\mu - \nu\|_{\text{TV}} \geq 0$ for all probability distributions μ, ν on Ω, and $\|\mu - \nu\|_{\text{TV}} = 0$ if and only if $\mu = \nu$.
- $\|\mu - \nu\|_{\text{TV}} = \|\nu - \mu\|_{\text{TV}}$
- $\|\mu - \nu\|_{\text{TV}} \leq \|\mu - \eta\|_{\text{TV}} + \|\eta - \nu\|_{\text{TV}}$ for all probability distributions μ, ν, η on Ω.

(Hint: you may want to use the triangle inequality for real numbers: $|x - y| \leq |x - z| + |z - y|$, $\forall\, x, y, z \in \mathbb{R}$.)

Exercise 9.71. Let μ, ν be probability distributions on a finite state space Ω. Then

$$\|\mu - \nu\|_{\text{TV}} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$$

(Hint: consider the set $A = \{x \in \Omega \colon \mu(x) \geq \nu(x)\}$.)

Theorem 9.72 (**The Convergence Theorem**)**.** *Let P be the transition matrix of a finite, irreducible, aperiodic Markov chain, with state space Ω and with (unique) stationary distribution π. Then there exist constants $\alpha \in (0,1)$ and $C > 0$ such that*

$$\max_{x \in \Omega} \|P^n(x, \cdot) - \pi(\cdot)\|_{\text{TV}} \leq C\alpha^n, \qquad \forall\, n \geq 1.$$

Proof. Since the Markov chain is irreducible and aperiodic, Lemma 9.67 implies there exists $r > 0$ such that all entries of P^r are positive. Let Π be the matrix with $|\Omega|$ rows, each of which is the row vector π (so $\Pi = (1, \ldots, 1)^T \pi$). From Theorem 9.43 (and Theorem 9.46), $\min_{z \in \Omega} \pi(z) > 0$. So, there exists $0 < \delta < 1$ such that

$$P^r(x, y) \geq \delta\pi(y), \qquad \forall\, x, y \in \Omega.$$

From Exercise 9.13, P^r is a stochastic matrix. Also, Π is a stochastic matrix. Let $\theta := 1 - \delta$. Define $Q := \theta^{-1}(P^r - (1 - \theta)\Pi)$. Then Q is a stochastic matrix, and

$$P^r = (1 - \theta)\Pi + \theta Q.$$

If M is an $|\Omega| \times |\Omega|$ stochastic matrix, then $M\Pi = \Pi$ (since $M\Pi = M(1, \ldots, 1)^T \pi = (1, \ldots, 1)^T \pi = \Pi$.) Similarly, if M satisfies $\pi M = \pi$, then $\Pi M = \Pi$. We now prove by induction that, for all $k \geq 1$,

$$P^{rk} = (1 - \theta^k)\Pi + \theta^k Q^k. \qquad (*)$$

We already know $k = 1$ holds, by the definition of Q. Assume $(*)$ holds for all $1 \leq k \leq n$. Then using $(*)$ twice,

$$\begin{aligned}
&P^{r(n+1)} = P^{rn} P^r = [(1 - \theta^n)\Pi + \theta^n Q^n] P^r \\
&= (1 - \theta^n)\Pi P^r + (1 - \theta)\theta^n Q^n \Pi + \theta^{n+1} Q^{n+1} \\
&= (1 - \theta^n)\Pi + (1 - \theta)\theta^n \Pi + \theta^{n+1} Q^{n+1}, \qquad \text{since } \pi P = \pi, \text{ so } \pi P^n = \pi, \text{ and } Q^n \text{ is stochastic} \\
&= (1 - \theta^{n+1})\Pi + \theta^{n+1} Q^{n+1}.
\end{aligned}$$

So, we have completed the inductive step, i.e. we have shown $(*)$ holds for all $k \geq 1$.

Let $j \geq 1$. Multiplying $(*)$ by P^j on the right and rearranging,

$$P^{rk+j} - \Pi = \theta^k (Q^k P^j - \Pi). \qquad (**)$$

From Exercise 9.13, $Q^k P^j$ is a stochastic matrix. Fix $x \in \Omega$. Sum up the absolute values of all the entries in row x of both sides of $(**)$ and divide by 2. By Exercise 9.71, the term on the right is then θ^k multiplied by the total variation distance between two probability distributions, which is at most 1, by definition of total variation distance. That is, the right side is at most θ^k. So, using Exercise 9.71 for the left side as well,

$$\left\| P^{rk+j}(x, \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} \leq \theta^k, \qquad \forall\, j, k \geq 1.$$

Taking the maximum of both sides over $x \in \Omega$, and writing an arbitrary positive integer n as $n = rk + j$ where $0 \leq j < r$ by Euclidean division of n by r (so that $k = (n/r) - (j/r) \geq (n/r) - 1$), we get the bound

$$\max_{x \in \Omega} \left\| P^n(x, \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} \leq \theta^{-1} (\theta^{1/r})^n.$$

Setting $C := \theta^{-1}$ and $\alpha := \theta^{1/r}$ completes the proof. □

9.8. **Markov Chain Monte Carlo.** We are now ready to describe the Markov Chain Monte Carlo method for simulating random variables. The **Markov Chain Monte Carlo** method constructs a Markov Chain with a (hopefully unique) stationary distribution that is equal to a given probability distribution. Simulation of the Markov chain itself then approximately simulates the given probability distribution. We introduce this topic by example.

Recalling the notation of Example 9.18, let $G = (V, E)$ be a finite graph. The **hard core model** μ is a probability measure on the set $\{0,1\}^V$. We can think of an element of $\xi \in \{0,1\}^V$ as particles lying on the graph G, so that $\xi(v) = 1$ if the vertex $v \in V$ contains a particle, and $\xi(v) = 0$ if the vertex $v \in V$ does not contain a particle. In this particle model, particles have repulsion, so that adjacent vertices cannot both contain particles. Let A denote the set of all elements of $\xi \in \{0,1\}^V$ such that, if $(v, w) \in E$ then $\xi(v), \xi(w)$ are not both equal to 1. Let z denote the number of elements of A. Then μ is defined to be the uniform distribution on A, i.e. $\mu(\xi) = 1/z$ for all $\xi \in A$, and $\mu(\xi) = 0$ for all $\xi \in \{0,1\}^V \setminus A$.

The measure μ we just defined is sufficiently complicated that, when G is a large graph, it is not easy to simulate μ using our previous simulation methods. Thankfully, there is an easy way to construct a Markov Chain $X_0, X_1, \ldots$ such that the distribution of X_n becomes close to that of μ as $n \to \infty$. This Markov Chain is constructed in the following way. Below, X_n will be a random element of $\{0,1\}^V$ for each $n \geq 0$. We also initialize X_0 to be the zero function on V. For any $n \geq 0$, we will define X_{n+1} using X_n.

For each integer $n \geq 0$, repeat the following procedure.

- Select one $v \in V$ uniformly at random.
- Let Y_n be uniformly distributed in $\{0,1\}$ and independent of all previously defined random variables.
- If $Y_n = 1$, and if all vertices $w \in V$ adjacent to v satisfy $X_n(w) = 0$, then set $X_{n+1}(v) := 1$. Otherwise, set $X_{n+1}(v) := 0$.
- For all $w \in V$ with $w \neq v$, define $X_{n+1}(w) := X_n(w)$.

Exercise 9.73. Show that the Markov Chain with state space A defined in our discussion of the hard core model is actually a Markov chain that is irreducible and aperiodic.

The Markov Chain $X_0, X_1, \ldots$ was created precisely because of the following fact.

Proposition 9.74. *The Markov chain $X_0, X_1, \ldots$ constructed above with state space A has unique stationary distribution μ.*

Proof. The above Markov chain $X_0, X_1, \ldots$ is irreducible by Exercise 9.73, so it has a unique stationary distribution π by Theorem 9.46. It remains to show that $\pi = \mu$. Since the stationary distribution is unique, it suffices to show that μ is stationary. It further suffices to show that μ is reversible, by Proposition 9.56. That is, it remains to show that

$$\mu(\xi)P(\xi,\zeta) = \mu(\zeta)P(\zeta,\xi), \qquad \forall\, \xi,\zeta \in A, \qquad (*)$$

where P is the transition matrix of the Markov chain P. Let d be the number of $v \in V$ such that $\xi(v) \neq \zeta(v)$. If $d = 0$, then $\xi = \zeta$, so both sides of $(*)$ are equal. If $d \geq 2$, then $P(\xi,\zeta) = P(\zeta,\xi) = 0$ by definition of the Markov chain, so both sides of $(*)$ are zero. To prove $(*)$ holds, it therefore remains to consider the case $d = 1$. If $d = 1$, then there exists exactly one vertex $v \in V$ such that $\xi(v) \neq \zeta(z)$, i.e. $\xi(w) = \zeta(w)$ for all other $w \in V$. Without loss of generality, assume $\xi(v) = 1$. Since $\xi \in A$, each neighbor w of v satisfies $\xi(w) = 0$, so that $\mu(\xi) = \mu(\zeta) > 0$ (recalling $\xi, \zeta \in A$), and $P(\xi,\zeta) = P(\zeta,\xi)$ since $\xi(w) = \zeta(w) = 0$ for all neighbors w of v, so $(*)$ holds. □

Remark 9.75. In fact, μ is stationary when the above Markov chain has state space $\{0,1\}^V$.

The Markov Chain we constructed above is a special case of a Gibbs Sampler where $S = \{0,1\}$.

Algorithm 9.76 (Gibbs Sampling Algorithm). Let S, V be finite sets. Let $\Omega := S^V$ be a state space. Let π be a probability distribution on Ω. The **Gibbs Sampling Algorithm** constructs a Markov Chain $X_0, X_1, \ldots$ with stationary distribution π. In this algorithm, X_n is a random element of S^V for each $n \geq 0$. We also initialize X_0 to be a constant function on V. For any $n \geq 0$, we will define X_{n+1} using X_n.

For each integer $n \geq 0$, repeat the following procedure.

- Select one $v \in V$ uniformly at random.
- Let $\overline{\pi}$ be the marginal distribution of π on v, given that all other vertex values are fixed by X_n. That is, $\overline{\pi}(s) := \pi\Big((X_n(w))_{w \in V \setminus \{v\}}, s\Big)$, $\forall\, s \in S$.
- Select $X_{n+1}(v)$ according to $\overline{\pi}$, i.e. $\mathbf{P}(X_{n+1}(v) = s) = \overline{\pi}(s)$, for all $s \in S$.
- For all $w \in V$ with $w \neq v$, define $X_{n+1}(w) := X_n(w)$.

Exercise 9.77. Show that the Gibbs Sampling Algorithm for a probability distribution π on Ω creates an aperiodic Markov Chain and π is reversible with respect to this Markov Chain. Conclude that π is a stationary distribution for the Markov Chain.

If it occurs that the Markov Chain is irreducible, conclude that π is the unique stationary distribution, so that the Gibbs Sampling Algorithm is a Markov Chain Monte Carlo Algorithm for simulating π.

Exercise 9.78. Let $G = (V,E)$ be a finite graph. Let $\lambda > 0$. The **generalized hard core model** μ_λ is a probability measure on the set $\{0,1\}^V$. Let A denote the set of all elements of $\xi \in \{0,1\}^V$ such that, if $(v,w) \in E$ then $\xi(v), \xi(w)$ are not both equal to 1. Then μ is defined so that $\mu(\xi) := 0$ for any $\xi \notin A$, and

$$\mu(\xi) := \frac{\lambda^{\sum_{v\in V}\xi(v)}}{z}, \qquad z := \sum_{\xi\in A} \lambda^{\sum_{v\in V}\xi(v)}.$$

- Show that, if $\xi \in \{0,1\}^V$, and if $v \in V$ is fixed with $\xi(w) = 0$ for all neighboring vertices w of v, then the probability that $\xi(v) = 1$ is $\lambda/(\lambda+1)$.
- Construct an MCMC algorithm for the generalized hard cord model.

The Metropolis Algorithm begins with a finite Markov Chain with transition matrix Q and state space Ω. Given a probability distribution π on Ω, the goal is to modify the Markov chain to obtain another Markov chain with stationary distribution π.

Algorithm 9.79 (**Metropolis Algorithm, Symmetric Case**). Suppose the matrix Q is symmetric. The **Metropolis Algorithm** constructs a Markov Chain $X_0, X_1, \ldots$ with stationary distribution π. The transition matrix P is defined so that

$$P(x,y) := \begin{cases} Q(x,y)\cdot \min\left(1,\ \frac{\pi(y)}{\pi(x)}\right) & \text{, if } y \neq x. \\ 1 - \sum_{z\in\Omega\colon z\neq x} Q(x,z)\cdot\min\left(1,\ \frac{\pi(z)}{\pi(x)}\right) & \text{, if } y = x. \end{cases}$$

To explain the choice of P, consider $a\colon \Omega\times\Omega \to [0,1]$. If $x \in \Omega$ is the current state of the original Markov chain and if $y \in \Omega$ is another state to move to, then $a(x,y)$ represents the probability that the state y is "accepted" by the new Markov Chain, and with remaining probability $1 - a(x,y)$ the chain is kept in its current state. With this intuition in mind, we create a transition matrix P of a more general form than above:

$$P(x,y) := \begin{cases} Q(x,y)a(x,y) & \text{, if } y \neq x. \\ 1 - \sum_{z\in\Omega\colon z\neq x} Q(x,z)a(x,z) & \text{, if } y = x. \end{cases}$$

In order for P to be reversible, we need

$$\pi(x)Q(x,y)a(x,y) = \pi(y)Q(y,x)a(y,x), \qquad \forall\, x,y\in\Omega,\ x\neq y.$$

Since Q is assumed to be symmetric, reversibility follows if and only if

$$\pi(x)a(x,y) = \pi(y)a(y,x), \qquad \forall\, x,y\in\Omega,\ x\neq y.$$

Since $a(x,y) \in [0,1]$, we must have the constraints

$$\pi(x)a(x,y) \leq \pi(x), \qquad \pi(x)a(x,y) = \pi(y)a(y,x) \leq \pi(y), \qquad \forall\, x,y\in\Omega,\ x\neq y.$$

We would like to choose values of $a(x,y)$ to be as large as possible, since rejecting moves from the original Markov Chain leads to slower simulations. The largest choice of a according to the above is $a(x,y) := \min(1, \pi(y)/\pi(x))$ for all $x, y \in \Omega$ with $x \neq y$.

Since P is reversible, recall that π is stationary for P by Proposition 9.56. However, the stationary distribution might not be unique.

Algorithm 9.80 (**Metropolis Algorithm, General Case**). The **Metropolis Algorithm** constructs a Markov Chain $X_0, X_1, \ldots$ with stationary distribution π. The transition matrix P is defined so that

$$P(x,y) := \begin{cases} Q(x,y)\cdot \min\left(1,\ \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right) & \text{, if } y \neq x. \\ 1 - \sum_{z\in\Omega\colon z\neq x} Q(x,z)\cdot\min\left(1,\ \frac{\pi(z)Q(z,x)}{\pi(x)Q(x,z)}\right) & \text{, if } y = x. \end{cases}$$

Exercise 9.81. Show that the transition matrix P constructed by the general case of the Metropolis Algorithm is reversible. Conclude that π is a stationary distribution for P.

Example 9.82. Suppose Q corresponds to the simple random walk on a large graph (such as a social network). In this case, Q might be easy to sample since neighbors of vertices are easy to find, but the global structure of the graph might be complicated. Suppose we want to sample from the uniform distribution on the graph. In general, this distribution will not be stationary for Q. In order to sample from the uniform distribution, we modify the simple random walk by using an acceptance probability $a(x,y):=\min(1,\deg(x)/\deg(y))$ for all $x\neq y$, $x,y\in V$. The Metropolis Algorithm constructs a Markov chain with this a with uniform stationary distribution. In words, this Markov Chain is biased against moving to high degree vertices.

Above we discussed how to construct Markov chains with a given stationary distribution π. An important question is then: given $\varepsilon>0$, what is the smallest integer $n>0$ such that $\max_{x\in\Omega}\|P^n(x,\cdot)-\pi(\cdot)\|_{\mathrm{TV}}<\varepsilon$? That is, how long does it take for the Markov Chain P to closely resemble the given distribution π? That is, what is the **mixing time** of the Markov chain P?

10. Appendix: Results from Analysis

Theorem 10.1 (**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 10.2. ***(Minkowski's Inequality)*** *Let $1\leq p\leq\infty$, and let $f\colon\mathbb{R}^2\to\mathbb{R}$ be measurable. Then*

$$\left\|\int_{\mathbb{R}} f(x,y)dx\right\|_{p,dy}\leq\int_{\mathbb{R}}\|f(x,y)\|_{p,dy}\,dx.$$

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 10.1. 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 1.71 (for Lebesgue measure), with exponents p,p' (using $(p-1)p'=p$).

$$\begin{aligned}\int_{\mathbb{R}}\left|\int_{\mathbb{R}} f(x,y)dx\right|^{p}dy&=\int_{\mathbb{R}}\left|\int_{\mathbb{R}} f(x,y)dx\right|^{p-1}\left|\int_{\mathbb{R}} f(x',y)dx'\right|dy\\&=\int_{\mathbb{R}}\Big(\int_{\mathbb{R}} f(x',y)\left|\int_{\mathbb{R}} f(x,y)dx\right|^{p-1}dy\Big)dx'\\&\leq\int_{\mathbb{R}}\Big(\int_{\mathbb{R}}|f(x',y)|^{p}\,dy\Big)^{1/p}\Big(\int_{\mathbb{R}}|\int_{\mathbb{R}} f(x,y)dx|^{p'(p-1)}dy\Big)^{1/p'}dx'\\&=\int_{\mathbb{R}}\|f(x',y)\|_{p,dy}\,dx'\cdot\Big(\int_{\mathbb{R}}|\int_{\mathbb{R}} f(x,y)dx|^{p}dy\Big)^{1/p'}.\end{aligned}$$

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 $\int_{\mathbb{R}} g(y)dy\leq 1$. Then by Fubini's Theorem and Hölder's inequality for y, Theorem 1.71 (for Lebesgue measure)

$$\int_{\mathbb{R}} g(y)\Big(\int_{\mathbb{R}} f(x,y)dx\Big)dy=\int_{\mathbb{R}}\Big(\int_{\mathbb{R}} f(x,y)g(y)dy\Big)dx\leq\int_{\mathbb{R}}\|f(x,y)\|_{\infty,dy}\,dx. \qquad (*)$$

From the Reverse Hölder inequality, if $h\colon\mathbb{R}\to\mathbb{R}$ is measurable, then

$$\|h\|_{\infty}=\sup_{\substack{g\colon\mathbb{R}\to[0,\infty)\\ \int_{\mathbb{R}} g(y)dy\leq 1}}\int_{\mathbb{R}} g(x)h(x)dx.$$

So, taking the supremum over such g in $(*)$, $\left\|\int_{\mathbb{R}} f(x,y)dx\right\|_{\infty,dy}\leq\int_{\mathbb{R}}\|f(x,y)\|_{\infty,dy}\,dx$. □

We say $f\colon\mathbb{R}\to\mathbb{R}$ is a **Schwartz function** if, for any integers $j,k\geq 1$, f is k times continuously differentiable and there exists $c_{j,k}\in\mathbb{R}$ such that

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

Proposition 10.3 (**Properties of Convolution on $\mathbb{R}$**). *Let $1\leq p\leq\infty$, 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 $1\leq p<\infty$, $\lim_{\varepsilon\downarrow 0}\|\phi_\varepsilon*f-cf\|_p=0$.*
(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) *$\forall\ m\geq 1$, $f*g$ is m times continuously differentiable, and $(f*g)^{(m)}=f^{(m)}*g$.*

Proof of (a),(b):

$$\begin{aligned}\|\phi_\varepsilon*f-cf\|_p&=\left\|\int_{\mathbb{R}}\phi_\varepsilon(y)(f(x-y)-f(x))dy\right\|_{p,dx}\\ &\leq\int_{\mathbb{R}}|\phi_\varepsilon(y)|\,\|f(x-y)-f(x)\|_{p,dx}\,dy\quad\text{, by Theorem. 10.2}\\ &=\int_{\mathbb{R}}|\phi(y)|\,\|f(x-\varepsilon y)-f(x)\|_{p,dx}\,dy,\text{ changing variables.}\end{aligned}$$

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 2.5, implies (a) and (b).

Proof of (c): Arguing as in (a) (taking absolute values, changing variables, and applying Dominated Convergence),

$$|(\phi_\varepsilon * f)(x) - cf(x)| \leq \int_{\mathbb{R}} |\phi(y)|\, |f(x-\varepsilon y) - f(x)|\, dy \to 0.$$

Proof of (d): Let $\eta > 0$. Choose $m > 0$ so that $2\left\|f\right\|_\infty \int_{|y|>m} |\phi(y)| \leq \eta$. Choose $\delta > 0$ by uniform continuity of f so that for any $x \in \mathbb{R}$, if $|u| \leq \delta$ then $|f(x+u) - f(x)| \leq \eta/\left\|\phi\right\|_1$. Then for any $0 < \varepsilon \leq \delta/m$ and for any $x \in \mathbb{R}$, if $|y| \leq m$, then $|f(x-\varepsilon y) - f(x)| \leq \eta/\left\|\phi\right\|_1$. So, continuing the calculation of (c), and applying the definition of m,

$$\int_{\mathbb{R}} |\phi(y)|\, |f(x-\varepsilon y) - f(x)|\, dy = \int_{\{y \in \mathbb{R}\colon\, |y|>m\}} (\cdots) + \int_{\{y \in \mathbb{R}\colon\, |y|\leq m\}} (\cdots)$$

$$\leq 2\left\|f\right\|_\infty \int_{\{y \in \mathbb{R}\colon\, |y|>m\}} |\phi(y)|\, dy + \int_{\{y \in \mathbb{R}\colon\, |y|\leq m\}} |\phi(y)| \frac{\eta}{\left\|\phi\right\|_1} \leq \eta + \eta = 2\eta.$$

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

$$\left|\frac{(f*g)(x+h) - (f*g)(x)}{h} - (f'*g)(x)\right| \leq \left\|\frac{f(x+h)-f(x)}{h} - f'(x)\right\|_{\infty,dx} \left\|g\right\|_1$$

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

Since f is a Schwartz function, $\left\|f''\right\|_\infty < \infty$, 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 10.4 (**Properties of Fourier Transform**). *Let f, g be Schwartz functions. Let $\xi \in \mathbb{R}$ and let $\lambda > 0$.*

(a) $|\widehat{f}(\xi)| \leq \int_{\mathbb{R}} |f(x)|\, dx$, $\forall\ \xi \in \mathbb{R}$.
(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): $|\widehat{f}(\xi)| = \left|\int_{\mathbb{R}} e^{ix\xi} f(x) dx\right| \leq \int_{\mathbb{R}} |f(x)|\, dx$.
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 10.1, and part (b) give

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

Proof of (e): Let $h > 0$. Using part (b) and the Dominated Convergence Theorem 2.5,

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

We now justify the use of the Dominated Convergence Theorem. By the Mean Value Theorem, $\left|\mathrm{Re}(e^{ixh}-1)/h\right| = |(\cos(xh)-1)/h| \leq |x|$ and $\left|\mathrm{Im}(e^{ixh}-1)/h\right| = |(\sin(xh)-1)/h| \leq |x|$, so $\left|(e^{ixh}-1)/h\right| \leq 2\,|x|$ and $\left|f(x)(e^{ixh}-1)/h\right| \leq 2\,|x|\,|f(x)|$.
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 10.1. □

Proposition 10.5. *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 10.4(a),(d) and Proposition 10.3(a),

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

as $\varepsilon \to 0$. Combining this statement with Proposition 10.4(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 10.4(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 10.4 shows that $\widehat{f}$ is k times differentiable for any $k \geq 1$, since f is a Schwartz function. And part (b) of the current Proposition says that $f^{(k)}$ vanishes at infinity for any $k \geq 1$, so repeated application of Proposition 10.4(f) shows that $\widehat{f}$ is a Schwartz function. □

Exercise 10.6. 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 10.7 (**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 10.4(c) and Proposition 10.5(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 10.4(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 10.4(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 2.5. And the right side tends to f uniformly in x by Proposition 10.3(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 10.8 (**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 10.9 (**Differentiating under the Integral Sign**). *Let $f\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$. Suppose*

- *For all $\theta \in \mathbb{R}$, $\int_{\mathbb{R}^n} |f(\theta, x)|\, dx < \infty$.*
- *For almost all $\theta \in \mathbb{R}$, the derivative $\partial f(\theta, x)/\partial\theta$ exists for all $x \in \mathbb{R}^n$.*

- *There is a function $g\colon \mathbb{R}^n \to [0,\infty)$ with $\int_{\mathbb{R}^n} |g(x)|\, dx < \infty$ and $|\partial f(\theta,x)/\partial\theta| \leq g(x)$ for all $\theta \in \mathbb{R}$, $x \in \mathbb{R}^n$.*

Then for all $\theta \in \mathbb{R}$,

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} f(\theta,x)dx = \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta} f(\theta,x)dx.$$

Proof. Let $h(\theta,x) := \frac{\partial}{\partial\theta} f(\theta,x)$ and let $h_0(\theta,x) := \int_0^\theta h(t,x)dt$ for any $\theta \in \mathbb{R}$, $x \in \mathbb{R}^n$. By assumption, $\int_{\mathbb{R}^n} |h(\theta,x)|\, dx < \infty$ for any $\theta \in \mathbb{R}$, so that $\int_0^\theta \int_{\mathbb{R}^n} |h(t,x)|\, dxdt < \infty$ for any $\theta \in \mathbb{R}$. By Fubini's Theorem 10.1,

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

Taking derivatives in θ of both sides and applying Lebesgue's Fundamental Theorem of Calculus, Theorem 10.10 (twice) concludes the proof. □

Theorem 10.10 (**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 $g'(x) = f(x)$.*

11. Appendix: Convergence in Distribution, Characteristic Functions

Definition 11.1 (**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 11.2. *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,

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

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

$$\begin{aligned}\mathbf{E}g(X) &= \mathbf{E}g(X)(1_{-2m<X\leq -m} + 1_{-m<X\leq t-\delta} + 1_{t-\delta<X\leq t})\\ &= O(\varepsilon) + F_X(t - \delta) + O(\varepsilon) = F_X(t) + O(\varepsilon).\end{aligned}$$

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,

$$\mathbf{P}(X_n \leq t) \geq \mathbf{E}g(X_n) \geq F_X(t) - O(\varepsilon), \qquad \forall\, n > n_0(\varepsilon).$$

Repeating the above with g where $g = 1$ on $(t + \delta, m]$ and $g = 0$ on $(-\infty, t] \cup [2m, \infty)$ gives

$$\mathbf{P}(X_n > t) \geq 1 - F_X(t) - O(\varepsilon), \qquad \forall\, n > n_0(\varepsilon).$$

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 11.3. *Let $\mu_1, \mu_2, \ldots$ be a sequence of probability measures on $\mathbb{R}$. Then any subsequential limit of the sequence (with respect to vague convergence) is a probability measure if and only if $\mu_1, \mu_2, \ldots$ is* ***tight****: $\forall\ \varepsilon > 0$, $\exists\ m = m(\varepsilon) > 0$ such that*

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

Exercise 11.4. Let $X, X_1, X_2, \ldots$ and let $Y, Y_1, Y_2, \ldots$ be random variables with values in $\mathbb{R}$.

(i) Assume that X is constant almost surely. Show that $X_1, X_2, \ldots$ converges to X in distribution if and only if $X_1, X_2, \ldots$ converges to X in probability.

(ii) Prove Lemma 11.3.

(iii) Suppose that $X_1, X_2, \ldots$ converges in distribution to X. Show there exist random variables $Z, Z_1, Z_2, \ldots\colon \Omega \to \mathbb{R}$ such that $\mu_Z = \mu_X$, $\mu_{Z_n} = \mu_{X_n}$ for any $n \geq 1$, and such that $Z_1, Z_2, \ldots$ converges almost surely to Z. (Hint: use Exercise 9.3.)

(iv) (Slutsky's Theorem) Suppose $X_1, X_2, \ldots$ converges in distribution to X and $Y_1, Y_2, \ldots$ converges in probability to Y. Assume Y is constant almost surely. Show that $X_1+Y_1, X_2+Y_2, \ldots$ converges in distribution to $X+Y$. Show also that $X_1Y_1, X_2Y_2, \ldots$

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 $X_1, X_2, \ldots$ converges in distribution to X, show that $\liminf_{n\to\infty} \mathbf{E}g(X_n) \geq \mathbf{E}g(X)$.

(vi) (Bounded convergence) If $g\colon \mathbb{R} \to \mathbb{C}$ is continuous and bounded, and if $X_1, X_2, \ldots$ converges in distribution to X, show that $\lim_{n\to\infty} \mathbf{E}g(X_n) = \mathbf{E}g(X)$.

(vii) (Dominated convergence) If $X_1, X_2, \ldots\colon \Omega \to \mathbb{R}$ converges in distribution to X, and if there exists a random variable $Y\colon \Omega \to [0,\infty)$ with $|X_n| \leq Y$ for all $n \geq 1$ and $\mathbf{E}Y < \infty$, show that $\lim_{n\to\infty} \mathbf{E}X_n = \mathbf{E}X$.

Theorem 11.5 (**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 11.4(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 10.7, 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 10.1,

$$\mathbf{E}g(X_n) = \frac{1}{2\pi}\int_{\mathbb{R}} \mathbf{E}e^{-iX_n y}\widehat{g}(y)dy = \frac{1}{2\pi}\int_{\mathbb{R}} \phi_{X_n}(-y)\widehat{g}(y)dy.$$

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 2.5 (and Proposition 10.5(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 10.3(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 11.2. □

Remark 11.6. 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 11.7 (**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 11.3 to get from (ii) to other conditions.)

12. Appendix: Moment Generating Functions

Exercise 12.1. 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 12.2. *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

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

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

then $|X_\varepsilon| \leq Y$ for any $0 < \varepsilon < \delta < 1$. We then conclude by the Dominated Convergence Theorem 2.5 that

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

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

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

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

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

Theorem 12.3 (**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 12.2 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 12.2 again, $(d/dt)^k|_{t=0}M_X(t) = \mathbf{E}X^k$ for all $k \geq 0$. 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 $0 < x < r$, $|x|^k \leq 1 + |x|^{k+1}$, so multiplying these inequalities and taking expected values gives

$$\frac{\mathbf{E}\,|X|^k\, r^k}{k!} \leq \frac{r^k}{k!} + \frac{\mathbf{E}\,|X|^{k+1}\, s^{k+1}}{(k+1)!}.$$

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,

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

We denote $\phi_X(t) := \mathbf{E}e^{itX}$. So, taking expected values of these same quantities with $x = X$,

$$\left|\phi_X(t+h) - \sum_{k=0}^{n} \frac{(i)^k \mathbf{E}e^{itX}X^k}{k!}\right| \leq \frac{|h|^{n+1}\,\mathbf{E}\,|X|^{n+1}}{(n+1)!}, \qquad \forall\, t \in \mathbb{R},\, \forall\, h \in (-\varepsilon, \varepsilon).$$

By $(**)$, the series then converges, so that

$$\phi_X(t+h) = \sum_{k=0}^{\infty} \frac{i^k \mathbf{E}e^{itX}X^k}{k!}h^k, \qquad \forall\, t \in \mathbb{R},\, \forall\, h \in (-\varepsilon, \varepsilon).$$

By Lemma 12.2, 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 11.6. □

13. 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 taking values in B. For example,
$f\colon \mathbb{R}^2 \to \mathbb{R}$ means that f is a function with domain $\mathbb{R}^2$ with values in $\mathbb{R}$

$\emptyset$ denotes the empty set

$A \subseteq B$ means $\forall\ a \in A$, we have $a \in B$, so A is contained in B

$A \smallsetminus B := \{a \in A\colon a \notin B\}$

$A^c := \Omega \smallsetminus A$, 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

$A \Delta B := (A \smallsetminus B) \cup (B \smallsetminus A)$

$\mathbf{P}$ 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

$\mathrm{var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2,$ the variance of X

$\sigma_X = \sqrt{\mathrm{var}(X)},$ 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 $t \in \mathbb{R}$.

$$(g * h)(t) = \int_{-\infty}^{\infty} g(x)h(t - x)dx \text{ denotes the convolution of } g \text{ and } h \text{ at } t \in \mathbb{R}$$

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