MATH 353 Probability and Statistics Summary

Benjamin T. Shepard

May 2, 2022

Executive Summary

This document is a summary of the topics covered in MATH 353 Probability and Statistics at Gettysburg College. It spans chapters 1 to 8 of the book *Introduction to Probability, Statistics and Random Processes* by Hossein Pishro-Nik. In particular, it covers the basic concepts of probability and its roots in set theory, combinatorial methods and their connections to probability, discrete and continuous random variables, joint distributions, expectation and variance, moment generating functions, the Law of Large Numbers, the Central Limit Theorem, and basic statistics such as point and interval estimators and hypothesis testing. Furthermore, it includes 1 to 2 fully worked solutions of problems given in the book for each chapter. These solutions are worked out in full detail, and only rely on the definitions and theorems that come before it in the summary. In other words, this document is self-contained. This project was excellent review for each topic that is covered, especially for the final exam. Furthermore, it was a great way to sharpen mathematical writing and LaTeX skills. It was also useful for tuning summary writing skills, as there is a lot of material that is covered in the book and condensing it into a short few pages is no easy task. Overall, this project was written initially in about 20 hours over the course of the Spring 2022 semester and was fine-tuned in the final week in a total of about 5 hours, making the total writing time around 25 hours. Although this document is already quite comprehensive, future improvements could be made to each section such as changing the order of definitions, adding more detailed introductions to each definition/theorem, adding dedicated worked examples to each chapter, and including additional solutions to worked problems. Finally, an in-depth improvement that would add to the overall quality of the document could be adding the proofs of the theorems that are presented in each summary, or at the very least, a sketch of each proof. This was not done here due to both the time constraint and the desire to avoid unnecessary complexity for the other students.

1 Basic Concepts

1.1 Chapter Summary

Consider rolling a die. Since beforehand we do not know the result, this is an example of a random experiment. This, along with set theory, is the basis for probability.

Definition 1.1 (Preliminary Terms). A *random experiment* is a process by which we observe something uncertain. An *outcome* is a result of a random experiment, and the set of all possible outcomes is the *sample space*. An *event* is a collection of possible outcomes.

Each event A is assigned a *probability measure* $\mathrm{P}(A)$, which outputs the likelihood that the event is observed. This is what we would like to eventually compute, but first, there are certain axioms that these measures must obey.

Definition 1.2 (Axioms of Probability). All probability measures must obey the following:

(1) For any event A, we have $\mathrm{P}(A) \geq 0$.

(2) The probability of the sample space S is $\mathrm{P}(S) = 1$.

(3) If $A_1, A_2, \ldots$ are disjoint events, then

$$\mathrm{P}\left(\bigcup_i A_i\right) = \sum_i \mathrm{P}(A_i).$$

In the case that the outcomes are all equally likely, finding the probability of a certain event is quite simple, as one might guess.

Theorem 1.3 (Equally Likely Outcomes). Let S be a finite sample space, and suppose that all outcomes are equally likely. Then for any event $A \subseteq S$, we have

$$\mathrm{P}(A) = \frac{|A|}{|S|}.$$

Another important concept is that of conditional probability. This lets us determine the likelihood of a certain event occurring given that another event has already occurred. We use this all the time in real life; for example, we might want to know the probability of rain given that there are clouds in the sky. Thus, we introduce the following notation.

Definition 1.4 (Conditional Probability). Let A and B be two events in some sample

space S. The *conditional probability of A given B*, as long as $\mathrm{P}(B) \neq 0$, is defined as

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

Perhaps the most important theorem to keep in mind when dealing with conditional probabilities is Bayes' rule. This rule gives us a way to "swap" the probability of the events; going back to our example, we might not know what the probability of rain given that there are clouds in the sky, but we could know the probability of clouds given that it is raining.

Theorem 1.5 (Bayes' Rule). If $B_1, B_2, \ldots$ partition the sample space S, and A is an event with $\mathrm{P}(A) \neq 0$, then we have

$$\mathrm{P}(B_j \mid A) = \frac{\mathrm{P}(A \mid B_j)\,\mathrm{P}(B_j)}{\sum_i \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i)}.$$

In particular, for any two events A and B with $\mathrm{P}(A) \neq 0$, we have

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \mid B)\,\mathrm{P}(B)}{\mathrm{P}(A)}.$$

The next important idea is that of independence. If two events do not have anything to do with one another, like getting a 1 when rolling a die and getting a 2 when rolling the die again, we say these events are independent; they do not affect each other at all.

Definition 1.6 (Independence). Two events A and B are *independent* if and only if

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B).$$

We can also combine the ideas of independence and conditional probability. It could be the case that two events are not independent to begin with, but if another event has occurred, they now become independent. For example, suppose there is a box with a regular coin and a two-headed coin. If I choose a coin at random, the events that I get heads on the first flip and I get heads on the second flip are not independent, because it might be the case that I chose the two-headed coin. However, they are independent *given* the condition that I chose the regular coin. The idea of conditional independence formalizes this.

Definition 1.7 (Conditional Independence). Two events A and B are *conditionally independent* given an event C, with $\mathrm{P}(C) \neq 0$, if

$$\mathrm{P}(A \cap B \mid C) = \mathrm{P}(A \mid C)\,\mathrm{P}(B \mid C).$$

Another important idea is the law of total probability, which lets us use a known partition of the sample space to our advantage. For example, if there is a known event that has a small probability of occurring (such as a device being defective) which affects another event's occurrence (like the lifespan of the device being less than 1 year), we can "separate" the two possibilities using the law of total probability, which makes calculating the probability of our desired event much easier.

Theorem 1.8 (Law of Total Probability). If $B_1, B_2, \ldots$ partition the sample space S, then for any event A we have

$$\mathrm{P}(A) = \sum_i \mathrm{P}(A \cap B_i) = \sum_i \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i).$$

1.2 Worked Problems

Problem 1.5.8. Let

$$A_n = \left[0, \frac{n-1}{n}\right)$$

for all $n \in \{2, 3, \ldots\}$. Find the quantity

$$A = \bigcup_{n=1}^{\infty} A_n.$$

Solution. We have $A = [0, 1)$.

Proof. First, let $x \in A$. Then $x \in A_n$ for some $n \in \{2, 3, \ldots\}$, i.e.

$$x \in \left[0, \frac{n-1}{n}\right).$$

Since $n \geq 2$, we have

$$\frac{n-1}{n} < 1$$

and thus

$$\left[0, \frac{n-1}{n}\right) \subset [0, 1)$$

which implies that $x \in [0, 1)$. Thus $A \subseteq [0, 1)$. Now, let $x \in [0, 1)$. Note that

$$\lim_{n \to \infty} \frac{n-1}{n} = 1.$$

Therefore, there is some $N \in \mathbb{N}$ large enough so that for all $n \geq N$, the quantity

$$\frac{n-1}{n}$$

gets arbitrarily close to 1. Therefore, there is some $k \geq N$ for which

$$x < \frac{k-1}{k} < 1.$$

This implies that

$$x \in \left[0, \frac{k-1}{k}\right) = A_k$$

so $x \in A$, as claimed. ■

Problem 1.5.15. I roll a fair die twice and obtain two numbers, X_1 and X_2, the result of the first and second roll, respectively.

(a) Find $\mathrm{P}(X_2 = 4)$.

(b) Find $\mathrm{P}(X_1 + X_2 = 7)$.

(c) Find $\mathrm{P}(X_1 \neq 2,\ X_2 \geq 4)$.

Solution.

(a) Since all outcomes are equally likely and each roll is independent of all others, we have

$$P(x_2 = 4) = \frac{1}{6}.$$

(b) The sample space after two rolls is

$$S = \{1, 2, 3, 4, 5, 6\}^2.$$

Let A be the event that $x_1 + x_2 = 7$. Then

$$A = \{(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)\}$$

which implies, using again the fact that all outcomes are equally likely,

$$P(A) = \frac{|A|}{|S|} = \frac{6}{36} = \frac{1}{6}.$$

(c) Let B be the event that $x_1 \neq 2$ and C be the event that $x_2 \geq 4$. Then

$$P(B) = 1 - P(x_1 = 2) = 1 - \frac{1}{6} = \frac{5}{6}.$$

Since

$$C = \{(4, x), (5, x), (6, x) \mid 1 \leq x \leq 6\}$$

we have $|C| = 18$, and

$$P(C) = \frac{|C|}{|S|} = \frac{18}{36} = \frac{1}{2}.$$

Thus, since each roll is independent of all others,

$$P(B \cap C) = P(B)P(C) = \frac{5}{6} \cdot \frac{1}{2} = \frac{5}{12}.$$

2 Combinatorics

2.1 Chapter Summary

Recall that for a finite sample space S with equally likely outcomes, we have

$$\mathrm{P}(A) = \frac{|A|}{|S|}$$

for all events A. Therefore, computing this probability comes down to a counting problem; that is, counting the number of times an element of A occurs. There are a few useful theorems that will help us with such problems. The first is known as the multiplication principle.

Theorem 2.1 (Multiplication Principle). Suppose that we perform $r \in \mathbb{N}$ experiments such that the kth experiment has n_k possible outcomes. Then there are a total of $n_1 \cdot \cdots \cdot n_r$ possible outcomes for the sequence of r experiments.

The next is useful when we wish to choose k objects from a set of n things, where the order matters and replacement is allowed. An example of this might be choosing an ordered pair (x, y) from a set of n numbers, where choices like (x, x) are allowed.

Theorem 2.2 (Ordered with Replacement). The number of ways to choose k objects from a set of $n \in \mathbb{N}$ elements is n^k.

The next is useful when we wish to choose k objects from a set of n things, where the order matters but replacement is not allowed. This is known as a *k-permutation of n elements*. An example of this might be choosing an ordered pair (x, y) from a set of n numbers, where choices like (x, x) are not allowed.

Theorem 2.3 (Ordered without Replacement). The number of ways to order k objects from a set of $n \in \mathbb{N}$ elements (i.e. the number of k-permutations of n elements), is

$$P_k^n = \frac{n!}{(n-k)!}.$$

The third is useful for choosing k elements from a set of n things, where the order does not matter and replacement is not allowed. This is known as a *k-combination of n elements*, and is represented with the choose function. An example of this might be choosing a 2-subset $\{x, y\}$ of a set of n things, where choices like $\{x, x\}$ are not allowed.

Theorem 2.4 (Unordered without Replacement). The number of k-combinations of $n \in \mathbb{N}$ elements is given by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

The final theorem is useful when we wish to choose k objects from a set of n things, where the order does not matter and replacement is allowed. An example of this might be choosing a way to write out the prime factorization of a number, since the order does not matter but there can be multiple of the same prime.

Theorem 2.5 (Unordered with Replacement). The number of unordered k-samples from $n \in \mathbb{N}$ elements is given by

$$\binom{n+k-1}{k}.$$

2.2 Worked Problems

Problem 2.2.2. Eight committee members are meeting in a room that has twelve chairs. In how many ways can they sit in the chairs?

Solution. Once a committee member chooses a seat, the others have one less option to choose from. Therefore, the order matters, and repetition is not allowed. Thus, the number of possibilities is the number of 8-permutations of 12 elements, or

$$P_8^{12} = \frac{12!}{4!}.$$

Problem 2.2.4. Five cards are dealt from a shuffled deck. What is the probability that the dealt hand contains the following?

(a) Exactly one ace.

(b) At least one ace.

Solution.

(a) There are 4 aces in the deck, so the number of ways to choose an ace is

$$\binom{4}{1},$$

and the number of ways to choose any four cards that are not aces is

$$\binom{48}{4}.$$

Thus the number of ways to choose exactly one ace out of 5 cards is

$$\binom{4}{1}\binom{48}{4}.$$

Since the number of ways to choose 5 cards from a deck of 52 is

$$\binom{52}{5},$$

the desired probability is

$$\frac{\binom{4}{1}\binom{48}{4}}{\binom{52}{5}}.$$

(b) The probability that the dealt hand contains at least one ace is 1 minus the probability that it contains no aces. Therefore, the desired probability is

$$1 - \frac{\binom{4}{0}\binom{48}{5}}{\binom{52}{5}}.$$

3 Discrete Random Variables

3.1 Chapter Summary

Random variables are a way to generalize the idea of focusing on a numerical aspect of a random experiment. For example, when rolling a die, we might want to know what the

probability of rolling a number less than 4 is. We can assign the numerical value of the roll to a random variable X, and then create an event A that X is less than 4. We then wish to find $\mathrm{P}(A)$. This example is very simple, but even in more complicated situations the concepts of random variables help us determine the likelihood of numerical events happening with ease.

Definition 3.1 (Random Variable). A *random variable* is a function $X : S \to \mathbb{R}$. Its range, denoted R_X, is the set of possible values of X.

Definition 3.2 (Discrete Random Variable). A random variable X is *discrete* if its range is countable; that is, there exists an injection $f : R_X \to \mathbb{N}$.

The probability mass function is a way to assign a probability to the random variable's value. In our previous example of rolling a die, we might want to know the probability that $X = 5$; using the PMF, this would be $\mathrm{P}_X(5)$.

Definition 3.3 (Probability Mass Function). Let X be a discrete random variable with range $R_X = \{x_1, x_2, \ldots\}$. The *probability mass function*, or PMF, of X is defined as

$$\mathrm{P}_X(x_k) = \mathrm{P}(X = x_k)$$

for all $k \in \mathbb{N}$.

Just like events, random variables can be independent of one another. This simplifies the situation greatly, so it is advantageous to recognize when this happens.

Definition 3.4 (Independence of Random Variables). Let X and Y be random variables. We say that X and Y are *independent* if

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}_X(x)\,\mathrm{P}_Y(y)$$

for all $x \in R_X$ and $y \in R_Y$.

There are a certain number of special distributions that random variables follow. Using these allow us to easily compute the PMF of a random variable. The first is a Bernoulli random variable. This distribution models random experiments that have two possible outcomes. An example of this would be a coin toss, which either lands heads or tails.

Definition 3.5 (Bernoulli Random Variable). A random variable X is called *Bernoulli*

with parameter p, where $p \in (0,1)$, written $X \sim \text{Bernoulli}(p)$, if its PMF is given by

$$\mathrm{P}_X(k) = \begin{cases} p & \text{if } k = 1; \\ 1-p & \text{if } k = 0; \\ 0 & \text{otherwise.} \end{cases}$$

The next is a geometric random variable. This distribution models the probability that k is the number of times a Bernoulli experiment is repeated until the first "success". An example of this would be flipping a coin until first observing a heads.

Definition 3.6 (Geometric Random Variable). A random variable X is called *Geometric with parameter* p, where $p \in (0,1)$, written $X \sim \text{Geometric}(p)$, if its PMF is given by

$$\mathrm{P}_X(k) = \begin{cases} p(1-p)^{k-1} & \text{if } k \in R_X; \\ 0 & \text{otherwise.} \end{cases}$$

The next is a binomial random variable. This distribution models the probability that k is the total number of successes observed over a given number of trials. An example of this would be the number of times a heads is observed when flipping a coin n times.

Definition 3.7 (Binomial Random Variable). A random variable X is called *Binomial with parameters* n *and* p, where $p \in (0,1)$, written $X \sim \text{Binomial}(n,p)$, if its PMF is given by

$$\mathrm{P}_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k \in \{0, \ldots, n\}; \\ 0 & \text{otherwise.} \end{cases}$$

The next is a Pascal random variable. This distribution models the probability that a Bernoulli experiment is repeated k times until observing m successes. This is a generalization of the geometric distribution, in which case $m = 1$. An example of this would be flipping a coin until observing heads m times.

Definition 3.8 (Pascal Random Variable). A random variable X is called *Binomial with parameters m and p*, where $p \in (0,1)$, written $X \sim \text{Pascal}(m,p)$, if its PMF is given by

$$\text{P}_X(k) = \begin{cases} \binom{k-1}{m-1} p^m (1-p)^{k-m} & \text{if } k \in \{m, m+1, \ldots\}; \\ 0 & \text{otherwise.} \end{cases}$$

The next is a hypergeometric random variable. This distribution models the probability of choosing k successes from a set with K elements. An example of this would be choosing k marbles out of a bag with b blue and r red marbles, where a success is getting a blue marble.

Definition 3.9 (Hypergeometric Random Variable). A random variable X is called *Hypergeometric with parameters b, r and p*, written $X \sim \text{Hypergeometric}(b,r,p)$, if its range is

$$R_X = \{\max\{0, k-r\}, \max\{0,k-r\}+1, \ldots, \min\{k,b\}\},$$

and its PMF is given by

$$\text{P}_X(x) = \begin{cases} \dfrac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}} & \text{if } x \in R_X; \\ 0 & \text{otherwise.} \end{cases}$$

The final distribution that random variables can follow is a Poisson random variable. This distribution models the probability of k events occurring in a given interval of time. An example of this would be the number of customers on average visiting a store within a given time period.

Definition 3.10 (Poisson Random Variable). A random variable X is called *Poisson with parameter λ*, written $X \sim \text{Poisson}(\lambda)$, if its range is $R_X = \mathbb{N}_0$ and its PMF is given by

$$\text{P}_X(k) = \begin{cases} \dfrac{\lambda^k}{e^k k!} & \text{if } k \in R_X; \\ 0 & \text{otherwise.} \end{cases}$$

There are also a few important functions and quantities related to random variables. One such function is the cumulative distribution function, which outputs the probability that the random variable is less than or equal to a given input. Its relation with the PMF is obvious for discrete variables, but for other types of random variables it is not so easily seen.

Definition 3.11 (Cumulative Distribution Function). The *cumulative distribution function*, or CDF, of X is defined as

$$F_X(x) = \mathrm{P}(X \leq x)$$

for all $x \in \mathbb{R}$.

Theorem 3.12 (Relation Between PMF and CDF). For all $a \leq b$, we have

$$\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a).$$

The next two quantities are the expectation and variance. These allow us to measure the average value of a random variable and how "spread out" its distribution is, respectively. The expectation measures the mean, and the variance measures the average level for which each data point is different from the mean.

Definition 3.13 (Expectation and Variance). Let X have range $R_X = \{x_1, x_2, \ldots\}$. The *expectation*, or *mean*, of X is defined as

$$\mathrm{E}[X] = \mu_X = \sum_{x_k \in R_X} x_k \, \mathrm{P}_X(x_k).$$

The *variance* of X is given by

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mu_X)^2] = \sum_{x_k \in R_X} (x_k - \mu_X)^2 \, \mathrm{P}_X(x_k) = \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

There are multiple properties that the expectation and variance must obey, which each make intuitive sense. The expectation is linear, while the variance is not – adding a constant to the random variable does not affect its value.

Theorem 3.14 (Linearity of Expectation). For all $a, b \in \mathbb{R}$, we have

$$\mathrm{E}[aX + b] = a \, \mathrm{E}[X] + b.$$

Theorem 3.15 (Nonlinearity of Variance). For all $a, b \in \mathbb{R}$, we have

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

Although the variance is nonlinear, in the case that there are multiple independent random variables, the variance acts linearly.

Theorem 3.16 (Variance of Independent Sum). If X can be written as

$$X = X_1 + \cdots + X_n$$

for some independent random variables $X_1, \ldots, X_n$, then

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n).$$

There is also the law of the unconscious statistician (LOTUS), which states that the expectation of a function of a random variable can just be brought inside the sum.

Theorem 3.17 (LOTUS). Let X be a discrete random variable. For any $g(X)$, we have

$$\mathrm{E}[g(X)] = \sum_{x_k \in R_X} g(x_k)\, \mathrm{P}_X(x_k).$$

Finally, the standard deviation is the square root of the variance. This is introduced in order to have a quantity similar to variance with the same unit as the random variable.

Definition 3.18 (Standard Deviation). The *standard deviation* of X is defined as

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

3.2 Worked Problems

Problem 3.3.2. Let X be the number of the cars being repaired at a repair shop. We have the following information:

- At any time, there are at most 3 cars being repaired.
- The probability of having 2 cars at the shop is equal to the probability of having 1 car.
- The probability of having no car at the shop is equal to the probability of having 3 cars.
- The probability of having 1 or 2 cars is half of the probability of having 0 or 3 cars.

Find the PMF of X.

Solution. The given information tells us that

- $0 \leq X \leq 3$,
- $\mathrm{P}_X(1) = \mathrm{P}_X(2)$,
- $\mathrm{P}_X(3) = \mathrm{P}_X(0)$, and

- $\mathrm{P}_X(1) + \mathrm{P}_X(2) = \frac{1}{2}\Big[\mathrm{P}_X(0) + \mathrm{P}_X(3)\Big]$

since $\mathrm{P}_X(a)$ and $\mathrm{P}_X(b)$ are disjoint for $a \neq b$. Now, the second and third items combine with the fourth to give us

$$2\,\mathrm{P}_X(1) = \frac{1}{2}\Big[2\,\mathrm{P}_X(0)\Big] \implies \mathrm{P}_X(0) = 2\,\mathrm{P}_X(1).$$

Thus, we have

$$\begin{aligned} 1 &= \sum_{k=0}^{3} \mathrm{P}_X(k) \\ &= \mathrm{P}_X(0) + \mathrm{P}_X(1) + \mathrm{P}_X(2) + \mathrm{P}_X(3) \\ &= 2\,\mathrm{P}_X(0) + 2\,\mathrm{P}_X(1) \\ &= 6\,\mathrm{P}_X(1). \end{aligned}$$

Therefore,

$$\mathrm{P}_X(1) = \mathrm{P}_X(2) = \frac{1}{6}$$

which implies that

$$\mathrm{P}_X(0) = \mathrm{P}_X(3) = \frac{2}{6} = \frac{1}{3}.$$

Thus, the PMF of X is

$$\mathrm{P}_X(k) = \begin{cases} 1/6 & \text{if } k = 1, 2; \\ 1/3 & \text{if } k = 0, 3; \\ 0 & \text{otherwise.} \end{cases}$$

Problem 3.3.11. The number of emails that I get in a weekday (Monday through Friday) can be modeled by a Poisson distribution with an average of $1/6$ emails per minute. The number of emails that I receive on weekends (Saturday and Sunday) can be modeled by a Poisson distribution with an average of $1/30$ emails per minute.

(a) What is the probability that I get no emails in an interval of length 4 hours on a Sunday?

(b) A random day is chosen, and a random interval of length one hour is selected on the chosen day. It is observed that I did not receive any emails in that interval. What is the probability that the chosen day is a weekday?

Solution.

(a) Let X be the number of emails recieved in the 4 hours. Then, since X is Poisson with

$$\lambda = 4 \cdot 60 \cdot \frac{1}{30} = 8,$$

we have

$$\mathrm{P}_X(0) = \frac{e^{-\lambda}\lambda^0}{0!} = \frac{1}{e^8}.$$

(b) Let Y be the number assigned to the day of the week (1 for Monday, 7 for Sunday). Then since all days are equally likely to be selected, we have

$$\mathrm{P}_Y(k) = \frac{1}{7}$$

for all $k \in \{1, \ldots, 7\}$. Using the fact that X is Poisson, its parameter for weekdays is

$$\lambda = 1 \cdot 60 \cdot \frac{1}{6} = 10$$

and for weekends is

$$\lambda = 1 \cdot 60 \cdot \frac{1}{30} = 2.$$

Thus,

$$\mathrm{P}(X = 0 \mid Y \leq 5) = \frac{1}{e^{10}} \quad \text{and} \quad \mathrm{P}(X = 0 \mid 6 \leq Y \leq 7) = \frac{1}{e^2}.$$

We also clearly have

$$\mathrm{P}(Y \leq 5) = \frac{5}{7} \quad \text{and} \quad \mathrm{P}(6 \leq Y \leq 7) = \frac{2}{7}.$$

Finally, by the law of total probability,

$$\begin{aligned}\mathrm{P}(X = 0) &= \mathrm{P}(X = 0 \mid Y \leq 5)\,\mathrm{P}(Y \leq 5) + \mathrm{P}(X = 0 \mid 6 \leq Y \leq 7)\,\mathrm{P}(6 \leq Y \leq 7) \\ &= \frac{1}{e^{10}} \cdot \frac{5}{7} + \frac{1}{e^2} \cdot \frac{2}{7} = \frac{5}{7e^{10}} + \frac{2}{7e^2}.\end{aligned}$$

Thus the desired probability is

$$\mathrm{P}(Y \leq 5 \,|\, X = 0) = \frac{\mathrm{P}(X = 0 \mid Y \leq 5)\,\mathrm{P}(Y \leq 5)}{\mathrm{P}(X = 0)} = \frac{\dfrac{5}{7e^{10}}}{\dfrac{5}{7e^{10}} + \dfrac{2}{7e^2}} \approx 0.0008.$$

4 Continuous Random Variables

4.1 Chapter Summary

Since discrete random variables can only take a countable number of possible values, we must develop new (yet similar) tools for dealing with random variables with uncountable ranges. Such random variables are called continuous.

Definition 4.1 (Continuous Random Variable). A random variable X is *continuous* if its range is uncountable.

Like discrete random variables, we can calculate the probability that a continuous random variable takes a certain value. However, since there are unaccountably many such values, we have $\mathrm{P}(X = x) = 0$ for all $x \in R_X$. We can still use the CDF to find the probability that a random variable is between two values, and we use this as the basis for the probability density function, which is the continuous analog of the PMF.

Definition 4.2 (Probability Density Function). Let X be a continuous random variable. The *probability density function*, or PDF, of X is defined as

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x),$$

as long as $F_X(x)$ is differentiable at x.

The rest of the definitions for continuous random variables mimic their respective counterparts in the discrete case. Recall that for a discrete random variable X, the expectation of X is given by

$$\mathrm{E}[X] = \sum_k k \, \mathrm{P}_X(k)$$

and the variance is given by

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mu)^2].$$

We define the expectation and variance for continuous random variables similarly.

Definition 4.3 (Expectation and Variance). Let X be a continuous random variable with integrable PDF $f_X(x)$. The expectation of X is defined as

$$\mathrm{E}[x] = \int_{-\infty}^{\infty} x f_X(x) \, \mathrm{d}x.$$

The variance of X is given by

$$\operatorname{Var}(X) = \mathrm{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,\mathrm{d}x.$$

The law of the unconscious statistician is also similar.

Theorem 4.4 (LOTUS). Let X be a continuous random variable. For any $g(X)$, we have

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x.$$

As we saw in the discrete case, there are a few distributions that continuous random variables can follow, the first of which is the most simple: the uniform distribution. This distribution describes a random experiment of an outcome being chosen lying in an interval. An example of this would be throwing a dart at a line on the wall.

Definition 4.5 (Uniform Distribution). A continuous random variable X is said to have *Uniform distribution over the interval* $[a,b]$, written $X \sim \operatorname{Uniform}(a,b)$, if its PDF is given by

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a < x < b; \\ 0 & \text{otherwise.} \end{cases}$$

The next is the exponential distribution. This distribution describes the time elapsed between events.

Definition 4.6 (Exponential Distribution). A continuous random variable X is said to have *exponential distribution with parameter* λ, written $X \sim \operatorname{Exponential}(\lambda)$, if its PDF is

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0; \\ 0 & \text{otherwise.} \end{cases}$$

The final distribution is known as the normal distribution. This describes random experiments that have a default mannerism and possible deviations from this mannerism. An example of this would be shooting an arrow at a target many times; usually, arrows are going to land near the bullseye, but there will be small deviations that make many arrows land away from the center. This is by far the most important distribution.

Definition 4.7 (Normal Random Variable). A continuous random variable X is said to be *normal with mean μ and variance σ^2*, written $X \sim N(\mu, \sigma^2)$, if its PDF is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

for all $x \in \mathbb{R}$. If $\mu = 0$ and $\sigma^2 = 1$, then X is called *standard normal*, written $X \sim N(0,1)$, and its PDF is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right].$$

4.2 Worked Problems

Problem 4.4.5. Let $X \sim \text{Uniform}(0,1)$ and let $Y = e^{-X}$.

(a) Find the CDF of Y.

(b) Find the PDF of Y.

(c) Find $\text{E}[Y]$.

Solution.

(a) The CDF of X for $x \in [0,2]$ is

$$F_X(x) = \frac{5}{32}\int_0^x u^4 \, \mathrm{d}u = \frac{u^5}{32}\bigg|_0^x = \frac{x^5}{32}.$$

Note that $R_Y = [0,4]$. Thus for $y \in [0,4]$, the CDF of Y is

$$F_Y(y) = \text{P}(Y \le y) = \text{P}(X^2 \le y) = \text{P}(X \le \sqrt{y}) = F_X(\sqrt{y}) = \frac{y^{5/2}}{32}.$$

Therefore,

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0; \\ \dfrac{y^{5/2}}{32} & \text{if } 0 \le y \le 4; \\ 1 & \text{if } y > 4. \end{cases}$$

(b) The PDF of Y for $y \in [0,4]$ is

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{5y^{3/2}}{64}.$$

Therefore,

$$f_Y(y) = \begin{cases} \dfrac{5y^{3/2}}{64} & \text{if } 0 \le y \le 4; \\ 0 & \text{otherwise.} \end{cases}$$

(c) The expectation of Y is

$$\mathrm{E}[Y] = \int_{R_Y} y f_Y(y)\,\mathrm{d}y = \frac{5}{64}\int_0^4 y^{5/2}\,\mathrm{d}y = \frac{10y^{7/2}}{448}\bigg|_0^4 = \frac{20}{7}.$$

5 Joint Distributions

5.1 Chapter Summary

If we wish to work with multiple random variables that are related to each other (i.e. dependent), we need to introduce new tools. We call random variables in this situation *joint*.

Definition 5.1 (Joint PMF). Let X and Y be two discrete random variables. The *joint probability mass function* of X and Y is defined as

$$\mathrm{P}_{XY}(x, y) = \mathrm{P}(X = x,\, Y = y).$$

The *joint range* of X and Y is then

$$R_{XY} = R_X \times R_Y = \{(x_i, y_j) \mid x_i \in R_X,\, y_j \in R_y\}.$$

The joint PMF gives us all of the information needed about the distributions of X and Y. We can use this to find the individual distributions through the marginal PMFs.

Definition 5.2 (Marginal PMF). The *marginal* PMFs of X and Y are defined as

$$\mathrm{P}_X(x) = \sum_{y_j \in R_Y} \mathrm{P}_{XY}(x, y_j) \quad \text{for any } x \in R_X$$

and

$$\mathrm{P}_Y(y) = \sum_{x_i \in R_X} \mathrm{P}_{XY}(x_i, y) \quad \text{for any } y \in R_Y.$$

Similar to the joint PMF, there is a joint CDF.

Definition 5.3 (Joint CDF). The *joint cumulative distribution function* of X and Y is

defined as

$$F_{XY}(x, y) = \mathrm{P}(X \leq x,\, Y \leq y).$$

When working with conditional probabilities, we need to define the conditional PMF and CDF as well.

Definition 5.4 (Conditional PMF). Let X and Y be discrete random variables and let A be an event. The *conditional PMF of X given A* is defined as

$$\mathrm{P}_{X|A}(x_i) = \mathrm{P}(X = x \mid A) = \frac{\mathrm{P}(X = x_i,\, A)}{\mathrm{P}(A)} \quad \text{for any } x_i \in R_X.$$

Furthermore, the *conditional PMF of X given Y* is defined as

$$\mathrm{P}_{X|Y}(x_i \mid y_i) = \frac{\mathrm{P}_{XY}(x_i, y_i)}{\mathrm{P}_Y(y_j)}$$

With this in mind, we can now define conditional expectation.

Definition 5.5 (Conditional Expectation). The *conditional expectation of X given A* is defined as

$$\mathrm{E}[X \mid A] = \sum_{x_i \in R_X} x_i \, \mathrm{P}_{X|A}(x_i).$$

Furthermore, the conditional expectation of X given that $Y = y$ is

$$\mathrm{E}[X \mid Y = y_j] = \sum_{x_i \in R_X} x_i \, \mathrm{P}_{X|Y}(x_i \mid y_j).$$

We can also define conditional variance.

Definition 5.6 (Conditional Variance). The *conditional variance of X given $Y = y$* is

$$\mathrm{Var}(X \mid Y = y) = \mathrm{E}[X^2 \mid Y = y] - \mathrm{E}[X \mid Y = y].$$

Now that we have seen how joint variables work, there is an alternate statement that gives us a condition for when two random variables are independent.

Theorem 5.7 (Independence). Two discrete random variables are independent if and only if

$$\mathrm{P}_{XY} = \mathrm{P}_X(x)\, \mathrm{P}_Y(y) \quad \text{for all } x, y.$$

It also lets us redefine LOTUS for discrete random variables.

Theorem 5.8 (LOTUS). Let X and Y be discrete random variables. For any function $g(X,Y)$, we have

$$\mathrm{E}[g(X,Y)] = \sum_{(x_i,y_j)\in R_{XY}} g(x_i,y_j)\,\mathrm{P}_{XY}(x_i,y_j).$$

Theorem 5.9 (Conditional Expectation Rules). If X and Y are independent random variables, then the following hold:

- $\mathrm{E}[X \mid Y] = \mathrm{E}[X]$,
- $\mathrm{E}[g(X) \mid Y] = \mathrm{E}[g(X)]$,
- $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$,
- $\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)]\,\mathrm{E}[h(Y)]$.

In order to determine how X and Y are related, we need to define covariance, which is a generalization of variance to two random variables.

Definition 5.10 (Covariance). The *covariance* of X and Y is defined as

$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y].$$

The covariance has the following properties.

Theorem 5.11 (Properties of Covariance). Let X, Y and Z be random variables. Then the following hold:

- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$,
- if X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$,
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$,
- $\mathrm{Cov}(aX,Y) = a\,\mathrm{Cov}(X,Y)$,
- $\mathrm{Cov}(X+c,Y) = \mathrm{Cov}(X,Y)$,
- $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$.

Covariance lets us determine how correlated two random variables are.

Definition 5.12 (Correlation Coefficient). The *correlation coefficient* of X and Y is defined as

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

Definition 5.13 (Correlated Variables). Let X and Y be random variables. Then

- if $\rho(X,Y) = 0$, then X and Y are *uncorrelated*,
- if $\rho(X,Y) > 0$, then X and Y are *positively correlated*,
- if $\rho(X,Y) < 0$, then X and Y are *negatively correlated*.

Theorem 5.14. If X and Y are independent, then they are uncorrelated and

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y).$$

5.2 Worked Problems

Problem 5.4.35. Let X and Y be two independent $N(0,1)$ random variables and let $Z = 7 + X + Y$ and $W = 1 + Y$. Find $\rho(Z,W)$.

Solution. Since $X, Y \sim N(0,1)$, we have

$$\mathrm{E}[X] = \mathrm{E}[Y] = 0$$

and

$$\operatorname{Var}(X) = \operatorname{Var}(Y) = 1.$$

Therefore, since X and Y are independent,

$$\begin{aligned}
\operatorname{Cov}(Z,W) &= \operatorname{Cov}(7 + X + Y, 1 + Y) \\
&= \operatorname{Cov}(X+Y, Y) \\
&= \operatorname{Cov}(Y,Y) + \operatorname{Cov}(X,Y) \\
&= \operatorname{Var}(Y) + 0 \\
&= 1.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\operatorname{Var}(Z) &= \operatorname{Var}(7 + X + Y) \\
&= \operatorname{Var}(X+Y) \\
&= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y) \\
&= 1 + 1 + 0 \\
&= 2
\end{aligned}$$

and

$$\mathrm{Var}(W) = \mathrm{Var}(1 + Y) = \mathrm{Var}(Y) = 1.$$

Therefore,

$$\rho(Z, W) = \frac{\mathrm{Cov}(Z, W)}{\sigma_Z \, \sigma_W} = \frac{1}{\sqrt{2}}.$$

6 Methods for More Than Two Random Variables

6.1 Chapter Summary

A generalization of expectation that gives us more information about the distribution of a random variable is the moment generating function.

Definition 6.1 (Moment). The nth moment of a random variable X is defined to be $\mathrm{E}[X^n]$.

Definition 6.2 (Moment Generating Function). The *moment generating function* of a random variable X is a function $M_X(t)$ defined as

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

This lets us get all moments of a random variable. In particular, we can obtain the expectation from the MGF, since it is the 1st moment, and the variance, since it is the 2nd moment minus the 1st moment squared.

Theorem 6.3 (Relation Between MGF and Moments). We have

$$M_X(t) = \sum_{k=0}^{\infty} \mathrm{E}[X^k] \, \frac{t^k}{k!}$$

and

$$\mathrm{E}[X^k] = \left(\frac{\mathrm{d}^k}{\mathrm{d}t^k} \, M_X(t) \right) \Big|_{t=0}.$$

Now we will define some bounds for the probability of random variables and events.

Theorem 6.4 (Union Bound). For any events $A_1, \ldots, A_n$, we have

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i \right) \leq \sum_{i=1}^{n} \mathrm{P}(A_i).$$

Theorem 6.5 (Markov's Inequality). If X is any nonnegative random variable, then for any $a > 0$ we have

$$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}[X]}{a}.$$

Theorem 6.6 (Chebyshev's Inequality). If X is any random variable, then for any $b > 0$ we have

$$\mathrm{P}(|X - \mathrm{E}[X]| \geq b) \leq \frac{\mathrm{Var}(X)}{b^2}.$$

6.2 Worked Problems

Problem 6.3.7. If

$$M_X(t) = \frac{1}{4} + \frac{1}{2}e^t + \frac{1}{4}e^{2t},$$

find $\mathrm{E}[X]$ and $\mathrm{Var}(X)$.

Solution. We can easily compute

$$\mathrm{E}[X] = M'(0) = \frac{1}{2}e^t + \frac{1}{2}e^{2t}\bigg|_0 = 1$$

and

$$\mathrm{E}[X^2] = M''(0) = \frac{1}{2}e^t + e^{2t}\bigg|_0 = \frac{3}{2}.$$

Therefore,

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = \frac{3}{2} - 1^2 = \frac{1}{2}.$$

Problem 6.3.22. The number of customers visiting a store during a day is a random variable with mean $\mathrm{E}[X] = 100$ and variance $\mathrm{Var}(X) = 225$.

(1) Find an upper bound for having more than 120 or less than 80 customers in a day.

(2) Find an upper bound for having more than 120 customers in a day.

Solution.

(1) Let X be the number of customers in a day. The probabilities of $X \leq 80$ and $X \geq 120$ are disjoint, so by Chebyshev's inequality,

$$\mathrm{P}(X \leq 80 \text{ or } X \geq 120) = \mathrm{P}(|X - 100| \geq 20) \geq \frac{\mathrm{Var}(X)}{20^2} = \frac{225}{400} = \frac{9}{16}.$$

(2) Using the one-sided Chebyshev inequality, we have

$$\mathrm{P}(X \geq 120) \leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + 120^2} = \frac{225}{225 + 120^2} = \frac{1}{65}.$$

7 Limit Theorems and Converence of Random Variables

7.1 Chapter Summary

The Law of Large Numbers has a very central role in probability and statistics. It states that if you repeat an experiment independently a large number of times and average the result, what you obtain should be close to the expected value. Before discussing the Law of Large Numbers, we need to first define the sample mean and iid random variables.

Definition 7.1 (IID Random Variables). Random variables $X_1, \ldots, X_n$ are said to be *iid* if they are independent and their distributions are identical.

Definition 7.2 (Sample Mean). For iid random variables $X_1, \ldots, X_n$, the sample mean, denoted by $\bar{X}$, is defined as

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$

Theorem 7.3 (Weak Law of Large Numbers). Let $X_1, \ldots, X_n$ be iid random variables with a finite expected value $\mathrm{E}[X_i] = \mu < \infty$. Then, for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathrm{P}(|\bar{X} - \mu| \geq \varepsilon) = 0.$$

The Central Limit Theorem (CLT) is one of the most important results in probability theory. It states that, under certain conditions, the sum of a large number of random variables is approximately normal.

Theorem 7.4 (Central Limit Theorem). Let $X_1, \ldots, X_n$ be iid random variables with finite expected value $\mathrm{E}[X_i = \mu < \infty$ and variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then the random variable

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

converges in distribution to the standard normal random variable as n grows large, that is,

$$\lim_{n\to\infty} \mathrm{P}(Z_n \le x) = \Phi(x)$$

for all $x \in \mathbb{R}$, where $\Phi(x)$ is the standard normal CDF.

7.2 Worked Problems

Problem 7.3.1. Let $X_i \sim \text{Uniform}(0,1)$ be iid.

(a) Find $\mathrm{E}[\bar{X}]$ and $\mathrm{Var}(\bar{X})$ as a function of n.

(b) Find an upper bound on

$$\mathrm{P}\left(\left|\bar{X} - \frac{1}{2}\right| \ge \frac{1}{100}\right).$$

(c) Using your bound, show that

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\bar{X} - \frac{1}{2}\right| \ge \frac{1}{100}\right) = 0.$$

Solution.

(a) Note that since $X_i \sim \text{Uniform}(0,1)$, we have $\mathrm{E}[X_i] = 1/2$ and $\mathrm{Var}(X_i) = 1/12$. By linearity of expectation,

$$\mathrm{E}[\bar{X}] = \frac{\mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n]}{n} = \frac{n\,\mathrm{E}[X_1]}{n} = \mathrm{E}[X_1] = \frac{1}{2}.$$

Also, since $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$ and the X_i's are independent,

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_1 + \cdots + X_n)}{n^2} = \frac{\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)}{n^2} = \frac{n\,\mathrm{Var}(X_1)}{n^2} = \frac{1}{12n}.$$

(b) By Chebyshev's inequality, we have

$$\mathrm{P}\left(\left|\bar{X} - \frac{1}{2}\right| \ge \frac{1}{100}\right) \le 100^2\,\mathrm{Var}(\bar{X}) = \frac{10000}{12n} = \frac{2500}{3n}.$$

(c) From part (b), and since probability is always non-negative, we have

$$0 \le \lim_{n\to\infty} \mathrm{P}\left(\left|\bar{X} - \frac{1}{2}\right| \ge \frac{1}{100}\right) \le \lim_{n\to\infty} \frac{2500}{3n} = 0$$

which implies the desired result by the squeeze theorem.

Problem 7.3.2. The number of accidents in a certain city is modeled by a Poisson random variable with an average rate of 10 accidents per day. Suppose that the number of accidents

on different days are independent. Use the Central Limit Theorem to find the probability that there will be more than 3800 accidents in a certain year.

Solution. Let X_i be the number of accidents that happen on day i. Then $X_i \sim \text{Poisson}(10)$, $\mathrm{E}[X_i] = \mathrm{Var}(X_i) = 10$, and the X_i's are independent. Then by the Central Limit Theorem, for any $n \in \mathbb{N}$,

$$\bar{X}_n \sim N\left(10, \frac{10}{n}\right).$$

This implies (where Z is a standard normal random variable)

$$\begin{aligned}
\mathrm{P}(X_1 + \cdots + X_{365} > 3800) &= \mathrm{P}\left(\frac{X_1 + \cdots + X_{365}}{365} > \frac{3800}{365}\right) \\
&= \mathrm{P}\left(\bar{X}_{365} > \frac{760}{73}\right) \\
&= \mathrm{P}\left(Z > \frac{760/73 - 10}{\sqrt{10/365}}\right) \\
&\approx \mathrm{P}(Z > 2.483) \\
&\approx 0.0065.
\end{aligned}$$

8 Statistical Inference I: Classical Methods

8.1 Chapter Summary

Here we wish to estimate a certain unknown parameter θ of a population. For example, θ might be the mean of a random variable, e.g. $\theta = \mathrm{E}[X]$. It is important to note that θ is not a random quantity, it is a fixed parameter.

Definition 8.1 (Point Estimator). Let θ be an unknown parameter of a population. Let $X_1, \ldots, X_n$ be a random sample. A *point estimator* for θ is a function of the random sample

$$\hat{\theta} = h(X_1, \ldots, X_n).$$

It is important to note that the estimator $\hat{\theta}$ *is* a random variable, while θ is not. When estimating θ, we will want to know how far off the estimator is from the actual value. This is called the bias of the estimator.

Definition 8.2 (Bias). Let $\hat{\theta}$ be a point estimator for θ. The *bias* of $\hat{\theta}$ is defined by

$$B(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta.$$

Definition 8.3 (Unbiased Estimator). Let $\hat{\theta}$ be a point estimator for θ. We say that $\hat{\theta}$ is an *unbiased* estimator for θ if $B(\hat{\theta}) = 0$ for all possible values of θ.

Note that if an estimator is unbiased, it is not necessarily a good estimator. Thus we introduce another measure to determine whether an estimator is good or not.

Definition 8.4 (Mean Squared Error). The *mean squared error* of a point estimator $\hat{\theta}$ is

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)^2].$$

Note that $\hat{\theta} - \theta$ is the error that we make when we estimate θ. Thus, the MSE is a measure of the distance between $\hat{\theta}$ and θ, and a smaller MSE is generally indicative of a better estimator. The last property that estimators have is consistency. Informally, an estimator is consistent if as the sample size gets larger, the estimator converges to the real value of the parameter.

Definition 8.5 (Consistent Estimator). Let $\hat{\theta}_1, \hat{\theta}_2, \ldots$ be a sequence of point estimators for θ. We say that $\hat{\theta}_n$ is a *consistent* estimator for θ if for all $\varepsilon > 0$ we have

$$\lim_{n \to \infty} \mathrm{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0.$$

Theorem 8.6 (Consistency Condition). Let $\hat{\theta}_1, \hat{\theta}_2, \ldots$ be a sequence of estimators for θ. If

$$\lim_{n \to \infty} \mathrm{MSE}(\hat{\theta}_n) = 0,$$

then $\hat{\theta}_n$ is a consistent estimator for θ.

We would like a systematic way of parameter estimation. To do this, we introduce an estimation method, called maximum likelihood estimation.

Definition 8.7 (Likelihood Function). Let $X_1, \ldots, X_n$ be a random sample from a distribution with parameter θ. Suppose that we have observed $X_1 = x_1, \ldots, X_n = x_n$. If the X_i's are discrete, the *likelihood function* is defined as

$$L(x_1, \ldots, x_n; \theta) = \mathrm{P}_{X_1 \cdots X_n}(x_1, \ldots, x_n; \theta) = \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n; \theta).$$

If the X_i's are jointly continuous, then the likelihood function is defined as

$$L(x_1, \ldots, x_n; \theta) = f_{X_1 \cdots X_n}(x_1, \ldots, x_n; \theta) = f(X_1 = x_1, \ldots, X_n = x_n; \theta).$$

Definition 8.8 (Maximum Likelihood Estimator). Let $X_1, \ldots, X_n$ be a random sample from a distribution with parameter θ. Suppose that we have observed $X_1 = x_1, \cdots, X_n = x_n$. The *maximum likelihood estimator* of θ is a random variable $\hat{\theta}_{ML}$ such that when $X_1 = x_1, \ldots, X_n = x_n$, its value is given by the value of θ that maximizes the likelihood function.

Now, we will introduce interval estimation. This gives us more information regarding how close the value of the estimator is to the actual value of the parameter. Instead of producing one point estimator, we produce an interval that hopefully includes the real value of the parameter.

Definition 8.9 (Interval Estimator). Let $X_1, \ldots, X_n$ be a random sample from a distribution with parameter θ. An *interval estimator* with *confidence interval* $1 - \alpha$ consists of two point estimators $\hat{\theta}_l(X_1, \ldots, X_n)$ and $\hat{\theta}_h(X_1, \ldots, X_n)$ such that

$$\mathrm{P}(\hat{\theta}_l \leq \theta \text{ and } \hat{\theta}_h \geq \theta) \geq 1 - \alpha$$

for every possible value of θ. Equivalently, we say that the interval $[\hat{\theta}_l, \hat{\theta}_h]$ is a $(1 - \alpha)$ 100% *confidence interval* for θ.

We will now introduce the concept of a hypothesis test. These are often used for comparing two hypotheses about a parameter of a population.

Definition 8.10 (Null & Alternative Hypothesis). Let θ be an unknown parameter, and let S be the set of all possible values for θ. Partition S into two disjoint sets S_0 and S_1. The *null hypothesis* is the statement that $\theta \in S_0$. The *alternative hypothesis* is that $\theta \in S_1$.

Most of the time, S_0 is a singleton; that is, the null hypothesis is that $\theta = \theta_0$ for some value θ_0, and the alternative is one of three choices: $\theta \neq \theta_0$, $\theta > \theta_0$, or $\theta < \theta_0$. Usually, the null hypothesis is the uninteresting statement, and the alternative is the one that we believe to be true. To decide between H_0 and H_1, we look at a function of the observed data, known as a test statistic. For example, the sample mean

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

is a statistic.

Definition 8.11 (Statistic). Let $X_1, \ldots, X_n$ be a random sample. A *statistic* is a real-valued function of the data. A *test statistic* is a statistic based on which we build our test.

Although useful, hypothesis tests are not perfect. When deciding between H_0 and H_1, there are two types of errors we can make.

Definition 8.12 (Acceptance Region). The *acceptance region* is defined to be the set $A \subset \mathbb{R}$ of possible values that the test statistic can take that would allow us to accept H_0. The *rejection region* is the set $R = \mathbb{R} \setminus A$.

Definition 8.13 (Type I/II Error). A *type I error* is the event that we reject H_0 when H_0 is true. A *type II error* is the event that we accept H_0 when H_0 is false.

Definition 8.14 (Significance Level). If the probability of a type I error satisfies

$$\mathrm{P}(\text{type I error}) \leq \alpha$$

for all $\theta \in S_0$, then we say the test has *significance level* α.

A related concept is the p-value, which is used as a cutoff to determine whether or not the null hypothesis should be rejected or not.

Definition 8.15 (P-value). The *p-value* is the lowest significance level α that results in rejecting the null hypothesis H_0.

For example, suppose that we set $\alpha = 0.05$. If we do our test assuming H_0 is true and the p-value is equal to 0.03, then we reject H_0. If the p-value instead is 0.07, then we cannot reject H_0, as the test has not demonstrated enough statistical significance for rejection. To conclude, we will define two procedures for hypothesis testing: for a proportion and the mean.

Example 8.16 (Hypothesis Test for a Proportion). We wish to test a claim about the proportion p of something in a population. The null and alternative hypotheses take the form

- $H_0 : p = p_0$
- $H_1 : p \neq p_0$ or $p < p_0$ or $p > p_0$

where p_0 is some fixed number. The test statistic in this case is called the z-test statistic and is computed using

$$z = \frac{\hat{p} - p_0}{\mathrm{SE}}$$

where $\hat{p}$ is the sample proportion from the random sample and

$$\mathrm{SE} = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

is the standard error. We then compute the p-value using this z-test statistic and the standard normal distribution table. If the p-value is less than the significance level α, we reject H_0. Otherwise, we fail to reject. The confidence interval is

$$(\hat{p} - \text{ME},\ \hat{p} + \text{ME})$$

where

$$\text{ME} = z \cdot \sqrt{\frac{\hat{p}(1-\hat{p}}{n}}$$

is the margin of error.

Example 8.17 (Hypothesis Test for the Mean). We wish to test a claim about the mean μ of a population. The null and alternative hypotheses take the form

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$ or $\mu < \mu_0$ or $\mu > \mu_0$

where μ_0 is some fixed number. The test statistic in this case is called the t-score and is computed using

$$t = \frac{\bar{X} - \mu_0}{\text{SE}}$$

where $\bar{X}$ is the sample mean from the random sample and

$$\text{SE} = \frac{S}{\sqrt{n}}$$

is the standard error, where S is the sample standard deviation. We then compute the p-value using this t-test statistic and a t-score distribution table. If the p-value is less than the significance level α, we reject H_0. Otherwise, we fail to reject. The confidence interval is

$$(\bar{X} - \text{ME},\ \bar{X} + \text{ME})$$

where $\text{ME} = t \cdot \text{SE}$ is the margin of error.

8.2 Worked Problems

Problem 8.6.3. Let $X_1, \ldots, X_n$ be a random sample from the following distribution:

$$f_X(x) = \begin{cases} \theta\left(x - \dfrac{1}{2}\right) + 1 & \text{for } 0 \leq x \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \in [-2, 2]$ is an unknown parameter. We define the estimator $\hat{\theta}_n$ as

$$\hat{\theta}_n = 12\bar{X} - 6.$$

(a) Is $\hat{\theta}_n$ an unbiased estimator of θ?

(b) Is $\hat{\theta}_n$ a consistent estimator of θ?

(c) Find the mean squared error of $\hat{\theta}_n$.

Solution.

(a) Note that for each $i \in \{1, \ldots, n\}$, we have

$$\begin{aligned}
\mathrm{E}[X_i] &= \int_0^1 x \left[\theta\left(x - \frac{1}{2}\right) + 1\right] \mathrm{d}x \\
&= \int_0^1 \left[\theta x^2 - \frac{\theta x}{2} + x\right] \mathrm{d}x \\
&= \frac{\theta x^3}{3} - \frac{\theta x^2}{4} + \frac{x^2}{2}\Bigg|_0^1 \\
&= \frac{\theta}{12} + \frac{1}{2}.
\end{aligned}$$

Therefore,

$$\mathrm{E}[\hat{\theta}_n] = \mathrm{E}[12\bar{X}_n - 6] = 12\,\mathrm{E}[\bar{X}_n] - 6 = 12\left(\frac{\theta}{12} + \frac{1}{2}\right) - 6 = \theta$$

so the bias of $\hat{\theta}_n$ is 0. Hence, $\hat{\theta}_n$ is an unbiased estimator of θ.

(b) For each $i \in \{1, \ldots, n\}$, we have

$$\begin{aligned}
\mathrm{E}[X_i^2] &= \int_0^1 x^2 \left[\theta\left(x - \frac{1}{2}\right) + 1\right] \mathrm{d}x \\
&= \int_0^1 \left[\theta x^3 - \frac{\theta x^2}{2} + x^2\right] \mathrm{d}x \\
&= \frac{\theta x^4}{4} - \frac{\theta x^3}{6} + \frac{x^3}{3}\Bigg|_0^1 \\
&= \frac{\theta}{12} + \frac{1}{3}.
\end{aligned}$$

Hence, since the X_i's are independent, we can use the linearity of expectation to get

$$\mathrm{E}[\hat{\theta}_n^2] = \mathrm{E}[(12\bar{X} - 6)^2]$$

$$
\begin{aligned}
&= 144\,\mathrm{E}[\bar{X}^2] - 144\,\mathrm{E}[\bar{X}] + 36 \\
&= \frac{144}{n^2}\,\mathrm{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - 144\,\mathrm{E}[\bar{X}] + 36 \\
&= \frac{144}{n^2}\,\mathrm{E}\left[\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n}\sum_{j\neq i}^{n} X_i X_j\right] - 144\,\mathrm{E}[\bar{X}] + 36 \\
&= \frac{144}{n^2}\left(n\,\mathrm{E}[X_i^2] + (n^2-n)\,\mathrm{E}[X_i]^2\right) - 144\,\mathrm{E}[\bar{X}] + 36 \\
&= \frac{144}{n^2}\left(n\left(\frac{\theta}{12}+\frac{1}{3}\right) + (n^2-n)\left(\frac{\theta}{12}+\frac{1}{2}\right)^2\right) - 144\left(\frac{\theta}{12}+\frac{1}{2}\right) + 36 \\
&= \theta^2 - \frac{\theta^2}{n} + \frac{12}{n}
\end{aligned}
$$

where the third from last line is obtained by noting that there are n^2 terms in total in the second sum, and n indexes are repeated. This gives us

$$\mathrm{Var}(\hat{\theta}_n) = \theta^2 - \frac{\theta^2}{n} + \frac{12}{n} - \theta^2 = \frac{12-\theta^2}{n}.$$

Let $\varepsilon > 0$. By Chebyshev's inequality,

$$0 \leq \lim_{n\to\infty} \mathrm{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{\mathrm{Var}(\hat{\theta}_n)}{\varepsilon^2} = \lim_{n\to\infty} \frac{12-\theta^2}{n\varepsilon^2} = 0.$$

Hence, by the squeeze theorem,

$$\lim_{n\to\infty} \mathrm{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0$$

so $\hat{\theta}_n$ is a consistent estimator of θ.

(c) By the above work, we have

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{Var}(\hat{\theta}_n) - B(\hat{\theta}_n)^2 = \frac{12-\theta^2}{n}.$$

Problem 8.6.3. Suppose we would like to test the hypothesis that at least 10% of students suffer from allergies. We collect a sample of 225 students and 21 of them suffer from allergies.

(a) State the null and alternative hypotheses.

(b) Obtain a test statistic and a p-value.

(c) State the conclusion at the $\alpha = 0.05$ level.

Solution.

(a) Let p be the proportion of students that suffer from allergies. The null is $H_0 : p \geq 0.1$. The alternative is $H_1 : p < 0.1$.

(b) Using the definition of standard error, we have

$$\text{SE} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.1(1-0.1)}{225}} = 0.02.$$

The sample proportion is

$$\hat{p} = \frac{21}{225} = 0.093,$$

so the z-test statistic becomes

$$z = \frac{\hat{p} - p_0}{\text{SE}} = \frac{0.093 - 0.1}{0.02} = -0.33.$$

Using a table, the corresponding p-value is 0.3707.

(c) If $\alpha = 0.05$, then the p-value is more than α. Therefore, we fail to reject H_0.