Chapter 2

Probability Theory

Experiments, Events and Probability

Shaikh Tanvir Hossain

October 26, 2022

East West University

Outline

1. Random Experiment

2. Probability - Definitions

3. Conditional Probability

Probability theory starts from Random Experiment. Here is the definition,

Definition 1.1 (Experiment and Event)

A *random experiment* is any process, real or hypothetical, in which *before performing* the experiment we can identify all possible outcomes but we don't know exactly which outcome will come.

- The *set* of all possible outcomes is called *sample space of the experiment*. We will use the notation ω to denote a single outcome and Ω to denote the sample space, this means $\Omega = \{\omega : \omega \text{ is an outcome of the experiment}\}$
- Any subset of the sample space is called an *event of the experiment*.

Note that the definition says before the experiment is performed we know all possible outcomes, but we do not know which outcome will come (so there is a lack of information or uncertainty!). Also another important thing, usually we can perform the same experiment more than once. When we perform the experiment a single time, we call it a *trial* of the experiment. Let's see some specific examples.

Here both Ω and ω are Greek letters, see `https://en.wikipedia.org/wiki/Omega`. This is pronounced as "Oh-may-gaa". Ω is the upper-case and ω is the lower-case

Here are some examples of Random Experiment.

- Tossing a coin. The sample space is $\Omega = \{\mathrm{H}, \mathrm{T}\}$
- Tossing two coins. The sample space is $\Omega = \{(\mathrm{H}, \mathrm{H}), (\mathrm{H}, \mathrm{T}), (\mathrm{T}, \mathrm{H}), (\mathrm{T}, \mathrm{T}\})$ (use multiplication rule to calculate the total number of possible outcomes)
- Throwing a die - The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Throwing two dice - The sample space is $\Omega = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 6), (2, 6), (2, 1), \ldots, (6, 6)\}$ (use multiplication rule to calculate the total number of possible outcomes, here total number of possible outcomes is 36.)

Another important example of random experiment is sampling,

- Sampling - The current population of Bangladesh is about $168,000,000$. Suppose we *randomly pick a sample* of 100 people so that it is a "good" representative of the population. This is a random experiment, because we don't know which 100 people will come in our sample, but we know the sample space Ω. It is the set of all people in Bangladesh. The sample in this case is called a *random sample*.
- It is important to note that in Statistics the bigger set from which we take our sample is called *population*. This may or may not mean literally population (জনসংখ্যা) of a country. This could be something else. It depends upon the what problem we are trying to solve.
- In Statistics we are often interested to know about the population, or some characteristics about the population (for example average income of the population) but what happens is we cannot access to the population, so try to get a random sample and then use that sample to say something about the population (we will see more about this later in our course!).
- We will come back to this. But for now just take the lesson that, *random sampling* is a very very important kind of random experiment. In fact most of the data that we analyze is a result of some kind of *random sampling*.

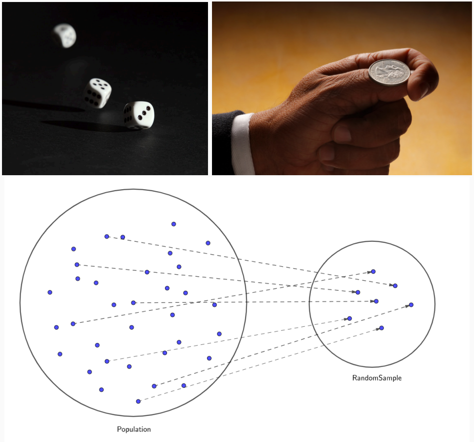


Figure 1: Throwing dices, tossing a single coin and sampling from a population, all are examples of random experiment!

- Once we know the sample space Ω, we can actually form different subsets of Ω, and think about different *events*. Recall an *events* is simply a subset of the sample space, so in principle everything that we have learned about Sets could be directly applicable when we are talking about Events.
- For example, if the sample space is $\Omega = \{H, T\}$, then
  - ✓ The set $\{H\}$ is an event since $\{H\} \subset \Omega$. Here event $\{H\}$ means only head is appearing.
  - ✓ Similarly the event $\{T\}$ means only tail is appearing.
  - ✓ Note $\{H, T\}$ is also an event, why?
  - ✓ Ques- What does the event $\{H, T\}$ mean?
  - ✓ Ques- What does the event $\emptyset$ mean?
- Can you think about *all possible events*? Yes, in this case, the answer is easy, we need the Power set of the sample space (remember, the power set is the set of all possible subsets)

$$\mathcal{P}(\Omega) = \{\{H\}, \{T\}, \Omega, \emptyset\}$$

- So each element in the set $\mathcal{P}(\Omega)$ is an event of the experiment. So in this case the Power set $\mathcal{P}(\Omega)$ is the set of all possible events.
- For an experiment the set of all possible events is also called the *Event Space*. Is the *Event Space* always going to be the Power Set? We will come back to events later again, now let's first define "what is a probability".

1. Random Experiment

2. **Probability - Definitions**

3. Conditional Probability

- Although all of us might have some intuitive understanding of probability, but the history of Mathematics tells us that the modern definition of probability came not so long ago.
- The Russian Mathematician Andrey Nikolaevich Kolmogorov (1903-87) laid the mathematical foundations of probability theory and the theory of randomness. His monograph *Grundbegriffe der Wahrscheinlichkeitsrechnung - Foundations of the Theory of Probability**, published in 1933 first introduced the Probability Theory in a rigorous way using fundamental axioms.
- We will see the *Axiomatic approach* of defining probability later, first let's see the *Classical approach* and *Frequentist approach* of defining probability.
- In all definitions we will calculate probability for events. For example if the set A is an event (i.e., $A \subset \Omega$ and $A \in \mathcal{P}(\Omega)$) then we will calculate $P(A)$, this is going to be a number in $[0, 1]$.

*Go to https://www.kolmogorov.com/Foundations.html to see the scanned version of the English translation.

Definition 1.1 - Classical Definition of Probability

If a Probability has n *equally likely* outcomes, and out of n, there are n_A outcomes which are associated with an event A, then the *probability of the event A* is, $\frac{n_A}{n}$. We will denote the probability of the event A by $P(A)$, this means

$$P(A) = \frac{n_A}{n} \tag{1}$$

- So when we are thinking about the event A, the classical definition says we can calculate the probability by,

$$P(A) = \frac{\text{number of outcomes in the event } A}{\text{total number of outcomes}}$$

- Let's apply the classical definition and calculate probabilities of some events of an experiment.

Example 1.2 (Applying the classical definition to calculate probabilities)

Suppose our Probability is throwing a *balanced die*. Note here *balanced die* means the outcomes *are all equally likely*. Here we have $\Omega = \{1, 2, 3, 4, 5, 6\}$, so $n = 6$. Let A be the event that *an even number occurs*. This means

$$A = \{2, 4, 6\}$$

We want to calculate $P(A)$. Here we have three outcomes for the event A (or associated with the event A), so $n_A = 3$, this means

$$P(A) = \frac{n_A}{n} = \frac{3}{6} = \frac{1}{2}$$

Example 1.3 (Applying the classical definition to calculate probabilities)

Suppose we toss 2 coins. Assume that all the outcomes are equally likely (fair coins).

- (a) What is the sample space?
- (b) Let A be the event that at least one of the coins shows up heads. Find $P(A)$.

Example 1.4 (Applying the classical definition to calculate probabilities)

Now suppose we toss 6 coins. Assume that all the outcomes are equally likely (fair coins).

- (a) How many elements are there in the sample space? Can you write one random element?
- (b) Let A be the event that we have heads in all 6 coins. Find $P(A)$.
- (c) Let B be the event that we have exactly one head and 5 tails. Find $P(B)$.

Example 1.5(Applying the classical definition to calculate probabilities)

Suppose in a city license plates have six characters: 3 letters followed by 3 numbers. Answer following questions,

- a) How many distinct such plates are possible?
- b) How many distinct plates are possible if the license plate contains no duplicate letters or numbers?
- c) Given that all sequences of six characters are equally likely, what is the probability that the license plate for a new car will contain no duplicate letters or numbers?

- a) We can apply multiplication rule, there are $26^3 = 17,576$ different ways to choose the letters and $10^3 = 1000$ ways to choose the numbers, so we have $17,576 \times 1000 = 17,576,000$ different plates. This means Ω consists of all $17,576,000$ possible sequences, so here $n = 17,576,000$
- b) Let's denote the event with A where we do not have any duplicates with numbers or digits. No duplicate letters means there are $26 \times 25 \times 24 = 15,600$ ways to choose the letters. And then there are $10 \times 9 \times 8 = 720$ ways to choose the numbers without duplication. From the multiplication principle, the number of outcomes in the event A is $15,600 \times 720 = 11,232,000$, so $n_A = 11,232,000$.
- c) So now we can calculate the probability of happening the event A,

$$P(A) = \frac{11,232,000}{17,576,000} = .64$$

- We will solve more examples in the practice sheet, now let's discuss the issues with the classical definition.
- There are essentially two major problems with the classical definition of probability
  - ✓ Assumption of equally likely outcomes (how do we know this?)
  - ✓ Finite sample space issues (sample space can be very large, e.g., $\Omega = \mathbb{R}$)

- Another definition is known as the *Frequency definition* of probability

Definition 1.6 - Frequency Definition of Probability

The probability of an event A is the relative proportion of outcomes if we perform the experiment *under identical condition* for a large number of times.

- So for example if our experiment is tossing a single coin, the probability of appearing heads is the number of times heads will appear if we perform this experiment almost infinite number of times.
- Frequency definition does not have equally likely outcomes assumption, but the issue is we need to perform the experiment *under identical conditions*, and this is often not possible.
- So in terms of the definition, the axiomatic definition does not have these issues, rather it's an abstract definition where we will define probability as a *set function*.

Definition1.7 - Axiomatic Definition of Probability

For a random experiment, if we have a sample space Ω and an associated event space $\mathcal{F}$, then we define probability as a *set function* P with domain $\mathcal{F}$ and codomain $\mathbb{R}_{\geq 0}$ (this means $P : \mathcal{F} \to \mathbb{R}_{\geq 0}$) that satisfies following 3 axioms,

- 1. **Axiom of Non-Negativity:** $P(A) \geq 0$ for all $A \in \mathcal{F}$.
- 2. **Axiom of Normalization:** $P(\Omega) = 1$.
- 3. **Axiom of Countable Additivity:** For a *sequence of pairwise disjoint events* $A_1, A_2, \ldots \in \mathcal{F}$ we have

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right)$$

- Let's explain each of these axioms (in class discussion).
- Note that, unlike the other definition, the Axiomatic definition does not tell us any ways to calculate probabilities, it only defines probability as a function.
- This means as long as any set function satisfies above three axioms, we will consider that function a probability function. Sometimes Probability function is also called *Probability measure*.

The idea of pairwise disjoint means, given that we have a sequence of sets $A_1, A_2, A_3, A_4, \ldots$, if we take any two sets then they won't have anything in common, this means for any i and j we will have $A_i \cap A_j = \emptyset$. We give the formal definitions later!

- With this definition, now we can show that the following rules of calculating probability

Theorem 1.8 (Probability Calculus)

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

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

Proof:

Let's do it on board...First we will prove c., then we will also see that b. holds and a. is also easy to show.

Here are the formal definitions of disjoint, pairwise disjoint and partition

Definition 1.9 (Disjoint, Pairwise Disjoint and Partition)

- Two events A and B are *disjoint* (or also called *mutually exclusive*) if $A \cap B = \emptyset$.
- The sequence of events $A_1, A_2, \ldots$ are *pairwise disjoint* (or *pairwise mutually exclusive*) if $A_i \cap A_j = \emptyset$ for any $i \neq j$.
- If $A_1, A_2, \ldots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = \Omega$, then the collection $A_1, A_2, \ldots$ forms a *partition* of Ω.

These concepts are easy to understand if we draw the Venn Diagrams.

Again just using the axioms, we can prove these results.

Theorem 1.10 (More Probability Calculus)

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

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

Again we can prove these claims using the definition. Using Ven Diagrams the ideas are easy, you will do this in your second problem set. But let's try to understand the theorem.

Recall we asked a question "*is the Power set of the sample space Ω always going to be the Event Space?*" The answer is NO! Not always. In particular when we have a very large sample space, for example $\Omega = \mathbb{R}$, the idea of Power Set might be problematic. But if *a family of sets* satisfies following three conditions, we can avoid pathological cases. First let's see the definition of an *Event Space*.

Definition 1.11 (Event Space)

A collection of subsets of Ω is called an *Event Space*, denoted by $\mathcal{F}$, if it satisfies the following three properties:

- a. $\emptyset \in \mathcal{F}$ (the empty set is an element of $\mathcal{F}$).
- b. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ ($\mathcal{F}$ is closed under complementation).
- c. If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ ($\mathcal{F}$ is closed under countable unions).

You should ask "why we need this?" The idea is If we do not put these conditions on the event space (or set of all possible events), we might get some bad sets as events where all three Axioms of Probability do not hold. In particular we might get negative probability of some sets. So Event space is definitely a subset of the Power set, but here we exclude the problematic sets. If a subset of power set satisfies above three conditions, it has a special name we call it *sigma-field*.

- Some important remarks from the last discussion is -
  - ✓ Not always Power set is a good option for set of events.
  - ✓ This is why there is a fixed definition about *what is a good set of events*, this is what we call *sigma-field*.
  - ✓ For any fixed samples space, it is possible to have many sigma-fields or event space. Usually we want it as big as possible. But it also depends on any particular context or problem.
- Notice with the axiomatic definition we have only defined Probability, but there is no rule or formula to calculate probabilities of any event.
- When we have a *finite sample space* then there is a nice rule to assign/calculate probability of an event A, following theorem gives us this rule.
- You have already applied this rule for the equally likely case. But now we don't need "equally likely assumption".

Theorem 1.12 (A rule to assign probabilities of events for a finite sample space)

Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be a finite sample space and $\mathcal{F}$ be *an Event space* of Ω. Let p_i be the probability that the outcome of the experiment will be ω_i and we have $p_1, \ldots, p_n$ such that following two conditions hold

1. $p_i \geq 0$ for all $i = 1, 2, \ldots, n$
2. $\sum_{i=1}^{n} p_i = 1$

Now for any $A \in \mathcal{F}$, we can define $P(A)$ by

$$P(A) = \sum_{\{i:\omega_i \in A\}} p_i$$

Also define $P(\emptyset) = 0$. Then we can show that P is a Probability function (this means Kolmogorov Axioms of Probability holds). This remains true if $\Omega = \{\omega_1, \omega_2, \ldots\}$ is a countable set.

Anderson et al. (2020) has the same rule in page 184. Note that Anderson et al. (2020) there is a slight abuse of notation, because it writes $P(E_i)$, in principle it should be $P(\{E_i\})$. This is abuse because in Anderson et al. (2020), E_i is an experimental outcome.

Let's see an example how to apply this theorem. Consider the simple experiment of tossing a coin, so $\Omega = \{\mathrm{H}, \mathrm{T}\}$. Now we don't need to assume the coin is "fair", so in this case maybe we can define

$$
\begin{aligned}
p_1 &= P(\{\mathrm{H}\}) = \frac{1}{9} \\
p_2 &= P(\{\ \mathrm{T}\}) = \frac{8}{9}
\end{aligned}
$$

Now, with this p_1 and p_2, we can calculate probabilities of any event A using the Theorem 1.12, and we can be sure according to the definition of probability this will be a Probability function. For example it is easy to check if $\mathcal{F} = \mathcal{P}(\Omega)$, then Axiom 1 and Axiom 3 holds. For checking Axiom 2 we need to check for every disjoint pairs.

Ques: What will be p_1 and p_2 if we assume "equally likely" experiment? You know the answer. What if our sample space is infinite. Actually we don't know we need to verify the axioms in each case, and this might be very hard! There is a solution if we consider random variables....(more on this later!)

1. Random Experiment

2. Probability - Definitions

3. **Conditional Probability**

- Now we will start with an important concept called *Conditioning*.
- All of the probabilities that we have dealt with thus far have been unconditional probabilities.
- A sample space was defined and all probabilities were calculated with respect to that sample space. In many instances, however, we are in a position to update the sample space based on new information.
- The idea is something has already happened and we want to update our probability calculation. This is known as *Conditional Probability*.
- Let’s see the definition first and then we will see some examples.

Definition 1.1 (Conditional Probability)

If A and B are events in Ω, and $P(B) > 0$, then the *conditional probability* of A given B, notation $P(A \mid B)$, is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{2}$$

- Note that what happens in the conditional probability calculation is that B becomes the sample space: $P(B \mid B) = 1$.
- The intuition is that our original sample space Ω has been updated to B. So All further calculations are then updated with respect to their relation to B.
- In particular, note what happens to conditional probabilities of disjoint sets. Suppose A and B are disjoint, so $P(A \cap B) = 0$. It then follows that $P(A \mid B) = P(B \mid A) = 0$. This means nothing will be updated for disjoint sets.
- Here $P(A \cap B)$ is sometimes called *joint probability*, because this is the probability of the joint event A and B.
- And $P(B)$ or $P(A)$, or probabilities for a single event are called *marginal probabilities*.

Example 1.2 (Conditional Probability)

So let's see an example, it is very common for patients with episodes of depression to have a recurrence (or relapse) within two to three years. Suppose we have studied 3 treatments for depression: *imipramine, lithium carbonate, and a combination*.

As is traditional in such studies (called clinical trials), there is also a group of patients who received a *Placebo*[†]

In this example, we shall consider 150 patients who entered the study. They were divided into the four groups (3 treatments and placebo) and followed to see how many had recurrences of depression. Following table summarizes the results.

	Treatment group				
Response	Imipramine	Lithium	Combination	Placebo	Total
Relapse	18	13	22	24	77
No relapse	22	25	16	10	73
Total	40	38	38	34	150

[†] A placebo is a treatment that is supposed to be neither helpful nor harmful. Some patients are given a placebo so that they will not know that they did not receive one of the other treatments. None of the other patients knew which treatment or placebo they received.

Here are couple of questions (For all questions assume, equally likely case, this means all patients have same probability of getting selected)

- What is the probability that a randomly selected patient had a relapse?
- What is the probability that a randomly selected patient received a placebo?
- What is the probability that a randomly selected patient received placebo and also had a relapse?
- Conditioning on the fact that a patient received placebo (or if we know that the patient received a placebo), what is the probability that the patient had a relapse?

Suppose A is the set of patients who had a relapse. Then calculating with the equally likely assumption, $P(A)$ can be calculated with

$$P(A) = \frac{\text{\# of patients who had relapse}}{\text{\# total patients}} = \frac{77}{150}$$

Now Let B be the event that the patient received placebo, then we can also calculate,

$$P(B) = \frac{\text{\# of patients who received placebo}}{\text{\# total patients}} = \frac{34}{150}$$

So we calculated the marginal probabilities of A and B. Now we can also calculate the joint probability $P(A \cap B)$. Where $A \cap B$ is the event where a randomly selected patient received a placebo and also had a relapse.

$$P(A \cap B) = \frac{24}{150}$$

With this, just applying the formula we can also calculate $P(A|B)$, which calculates the conditional probability,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{24/150}{34/150} = \frac{24}{34}$$

Note that there is a difference between the event $A \cap B$ and event we are considering when we calculate $P(A|B)$.

- The conditional probability definition gives a way to calculate the probabilities of a joint event. Note that using the formula in (2) we can easily get.

$$P(A \cap B) = P(A|B)P(B) \tag{3}$$

- This is sometimes called *the multiplication rule of conditional probability* (do not confuse this with the multiplication rule for counting!)
- Now using the same idea in (2) we can also calculate

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \text{ given that } P(A) \neq 0 \tag{4}$$

- From here using the multiplication idea, we get

$$P(A \cap B) = P(B|A)P(A)$$

- So now we have a different way of writing $P(A|B)$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{5}$$

- The last formula where we "turned around" the conditional probabilities is called *Bayes' Rule*, this is after the name of Sir Thomas Bayes.

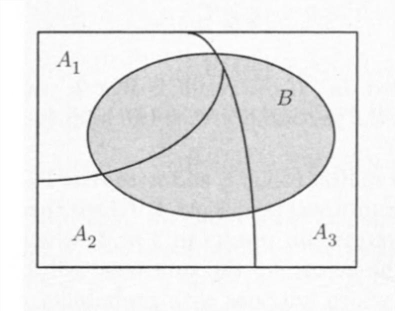
- Bayes' rule has a more general version (we will see this in a minute). But to understand this we need to first understand the *Law of Total Probability*
- Law of total Probability is an application of Partition, here is what the law says,

Theorem 1.3 (Law of Total Probability)

Let $A_1, \ldots, A_n$ be events that form a partition of the sample space Ω and assume that $P(A_i) > 0$, for all i. Then, for any event B, we have

$$P(B) = P(A_1)\, P(B \mid A_1) + \cdots + P(A_n)\, P(B \mid A_n)$$

We will do an informal visual proof. Visually this is very easy to understand,



First let's see for three sets, using the idea in the figure note that

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B)$$

All sets are disjoint, so using Axiom 3,

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

Since, by the definition of conditional probability, we have

$$P\left(A_i \cap B\right) = P\left(A_i\right) P\left(B \mid A_i\right) \text{ for } i = 1, 2, 3$$

Now we get

$$\begin{aligned} P(B) &= P\left(A_1 \cap B\right) + P\left(A_2 \cap B\right) + P\left(A_3 \cap B\right) \\ &= P\left(A_1\right) P\left(B \mid A_1\right) + P\left(A_2\right) P\left(B \mid A_2\right) + P\left(A_3\right) P\left(B \mid A_3\right) \end{aligned}$$

Now we can extend this idea for n dosjoint sets,

$$B = (A_1 \cap B) \cup \cdots \cup (A_n \cap B)$$

Again, using Axiom - 3. it follows that

$$P(B) = P(A_1 \cap B) + \cdots + P(A_n \cap B)$$

Since, by the definition of conditional probability, we have

$$P(A_i \cap B) = P(A_i) P(B \mid A_I).$$

And we get

$$\begin{aligned} P(B) &= P(A_1 \cap B) + \cdots + P(A_n \cap B) \\ &= P(A_1) P(B \mid A_1) + \cdots + P(A_n) P(B \mid A_n) \end{aligned}$$

Now we have the general version of the Bayes Rule

Theorem 1.4 (Bayes' Rule)

Let $A_1, A_2, \ldots, A_n$ be events that form a partition of the sample space Ω, and assume that $P(A_i) > 0$, for all i. Then, for any event B such that $P(B) > 0$, we have

$$\begin{aligned} P(A_i \mid B) &= \frac{P(A_i) P(B \mid A_i)}{P(B)} \\ &= \frac{P(A_i) P(B \mid A_i)}{P(A_1) P(B \mid A_1) + \cdots + P(A_n) P(B \mid A_n)} \end{aligned}$$

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., Cochran, J. J., Fry, M. J., and Ohlmann, J. W. (2020). *Statistics for Business & Economics*. Cengage, Boston, MA, 14th edition.

Bertsekas, D. and Tsitsiklis, J. N. (2008). *Introduction to probability*. Athena Scientific, 2nd edition.

Blitzstein, J. K. and Hwang, J. (2015). *Introduction to Probability*.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Thomson Learning, Australia ; Pacific Grove, CA, 2nd edition.

DeGroot, M. H. and Schervish, M. J. (2012). *Probability and Statistics*. Addison-Wesley, Boston, 4th edition.

Hansen, B. (2022). *Econometrics*. Princeton University Press, Princeton.

Newbold, P., Carlson, W. L., and Thorne, B. M. (2020). *Statistics for Business and Economics*. Pearson, Harlow, England, 9th, global edition.

Pishro-Nik, H. (2016). *Introduction to probability, statistics, and random processes*.

Ramachandran, K. M. and Tsokos, C. P. (2020). *Mathematical Statistics with Applications in R*. Elsevier, Philadelphia, 3rd edition.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*. Duxbury advanced series. Thomson/Brooks/Cole, Belmont, CA, 3rd edition.