Basic Probability Theory

Dustin Pluta

Graduate Statistics Bootcamp
Department of Statistics
University of California, Irvine

September 9th, 2019

Overview

- Welcome to UCI Statistics!
- Goal of these eight days is to teach you the basics of probability and statistics, review mathematical concepts, and teach basic programming.
- The course will serve as a review of important concepts and a preview for what you'll be learning through most of the year
- **Dates:** 9/9 - 9/13, 9/16 - 9/18
- **Time:** 9am - 4pm
- **Room:** ICS-1 432
- **Website:** `github.com/dspluta/StatsGradBootcamp2019`

Overview of Topics

1) Basic Probability
2) Calculus Review
3) Intro to R
4) Probability Distributions
5) Linear Algebra Review
6) Collections of Random Variables
7) Estimation and Inference
8) Intro to Statistical Modeling
9) Statistical Computing in R

Programming

- Download R & RStudio if you don't have them.
  - Download R - `https://cran.r-project.org/`
  - Download RStudio - `https://www.rstudio.com/products/rstudio/download3/`
- If you are new to R, the `swirl` package provides a good introduction. The swirl website provides Step-by-step instructions for getting started: `swirlstats.com/students.html`.

Overview of Books

These are some good reference texts to get you started.

- Undergraduate Statistics
  - *Mind on Statistics* - Jessica Utts & Robert Heckard
  - *A First Course in Probability* - Sheldon Ross
  - *Probability and Statistics* - Morris DeGroot & Mark Schervisch
  - *Applied Linear Regression Models* - Kutner, Nachtscheim, Neter
- Graduate Statistics
  - *Statistical Inference* - George Casella & Roger Berger
  - *Modern Mathematical Statistics* - Edward Dudewicz & Satya Mishra
  - *All of Statistics: A Concise Course in Statistical Inference* - Larry Wasserman

Brief Overview of UCI Statistics PhD Program

Degree Requirements:
`www.stat.uci.edu/m-s-ph-d-in-statistics/`
Required Courses:

1. 1st Year Theory: 200 A/B/C (Probability & Inference)
2. 1st Year Methods: 210/211/212 (Linear Models)
3. 2nd Year Theory: 220 A/B (Measure Theory and Asymptotics)
4. 2nd Year Methods: 225 (Bayesian Analysis) & 230 (Statistical Computing)
5. 275 (Statistical Consulting)
6. Four additional graduate courses in or related to statistics.
7. Enrollment in STAT 280 in all quarters.

Old qualifying exams are available at `www.stat.uci.edu/grad/exam/`.

Brief Overview of UCI Statistics PhD Program

Additional Requirements:

1. 1st Year Exams on Theory (200 A/B/C) and Methods (210/211/212).
2. 1st Year data analysis exam.
3. 2nd year Exams on Theory (220 A/B).
4. **Advancement Exam**, typically in the third year.
5. **Dissertation and Final Defense**.

UCI Statistics PhD Tips

1. Keep showing up!
2. In the first two years, focus on passing the core classes and the exams.
3. Work together, help each other learn the material and study for the exams.
4. Talk to professors and other students about their research.
5. Practice programming and good software design.

Thinking about Statistics

- What is Statistics?
- What kinds of problems and questions is statistics good for?
- What are some of the limitations and challenges in the application of statistics today?
- How is the landscape of statistics changing in response to trends in machine learning and data science?

What is Statistics?

- "[A] branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data." – Merriam-Webster
- "Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data.... Two fundamental ideas in the field of statistics are uncertainty and variation. " – UCI Stats
- "[S]tatistics is concerned with the use of data in the context of uncertainty and decision making in the face of uncertainty." – Wikipedia
- "Statistics is a mathematical framework for synthesizing prior knowledge, beliefs, and observational data to draw conclusions in the presence of uncertainty." – D. Pluta

Applications of Statistics

- Estimating the effect of a new drug therapy in reducing risk of heart attack.
- Identifying genes that are significantly associated with Alzheimer's disease.
- Predicting air quality in Orange County over time. `https://air.plumelabs.com/en/live/los-angeles`
- Classifying unlabeled photos based on the image scene.
- `https://projects.fivethirtyeight.com/2018-midterm-election-forecast/house/?ex_cid=rrpromo`

The Challenge of Statistics

- Statistical analysis can only be good as the data...
- Only certain questions can be answered with a given data set. Statistics is (in part) about making the best use of the data at hand to answer the questions of interest.
- "The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data" – John Tukey
- "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of." – R. A. Fisher

Statistics Research

- Survival analysis
- Epidemiology
- Computational statistics
- Bayesian Methods
- Nonparametric statistics
- Time series and forecasting
- Survey sampling

What is Probability?

- Informal definition (Wikipedia): Probability is the measure of the likelihood that an event will occur.
  - Loaded with terms... *measure, likelihood, event*...
- Probability can be thought of as the "frequency" with which a certain event would occur given a specific system (Frequentism)
- Probability can also be a way to quantify our beliefs about the world (Bayesian)
- Statistics is built around the language of probability

Beginning to think about Probability - Sample Spaces & Events

- Want to refamiliarize everyone with the foundations of probability
- Consider a system, or experiment, that we are interested in understanding it's apparently random behavior
- The system or experiment usually has some *set* of potential outcomes or results which could occur
- The **sample space** will be the set of all possible outcome **events**

Formal Definitions - Sample Spaces & Events

- We can begin to formalize some of the language that we use to talk about probability

Definition (Sample Space)

The set Ω of all possible outcomes of a particular experiment is called the sample space for the experiment.

Definition (Event)

An event is any collection of possible outcomes of an experiment, that is, any subset of Ω (including Ω itself).

Example - Sample Spaces & Events

- Tossing a coin - Each side of the coin is a potential outcome of the experiment, thus the sample space can written as follows

$$\Omega = \{\text{`Heads'}, \text{`Tails'}\}$$

- Rolling a dice - Each face of the die is a potential outcome of the experiment, thus

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

- Sum of the faces of two rolled dice - Worst case is two ones, best case is two sixes

$$\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Reviewing Set Theory I

- Since sample spaces are sets, it is worthwhile talking about operations of sets.
- Consider two events E, F in Ω

Definition (Union)

The union of E and F, written $E \cup F$, is the set of elements that belong to E or F or both.

Definition (Intersection)

The intersection E and F, written $E \cap F$, is the set of elements that belong to E and F.

Reviewing Set Theory II

Definition (Complement)

The complement of E written E^c is the set of elements not in E

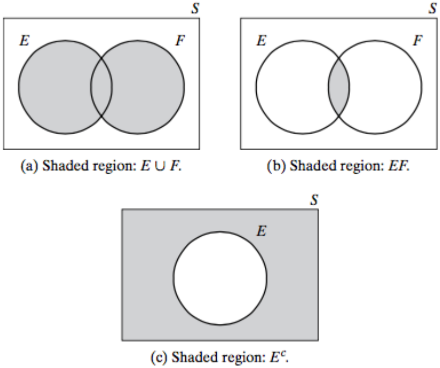


FIGURE 2.1: Venn Diagrams

Reviewing Set Theory III

Definition (Mutually Exclusive)

Two events E and F are mutually exclusive if their intersection is the empty set, that is, $E \cap F = \emptyset$.

- We can define unions and intersections on more than two sets similarly. If $E_1, E_2, \cdots \in \Omega$, then their union can be written $\cup_{i=1}^{\infty} E_i$ for $n = 1, 2, \ldots$.
- Similarly, if $E_1, E_2, \cdots \in \Omega$, then their intersection can be written $\cap_{i=1}^{\infty} E_i$ for $n = 1, 2, \ldots$

Reviewing Set Theory IV

- Finally we can discuss ideas of equality and containment of sets.

Definition (Subset)

For two events E and F, if all of the outcomes in E are also in F then we say that E is contained in F, written $E \subset F$, or E is a subset of F.

- The concept of equality can be developed such that if $E \subset F$ and $F \subset E$, then $E = F$.

Operations on Sets I

We can also talk about the rules and operations regarding sets

- Commutative Laws

$$E \cup F = F \cup E \quad \text{and} \quad E \cap F = F \cap E$$

- Associative Laws

$$(E \cup F) \cup G = E \cup (F \cup G) \quad \text{and}$$
$$(E \cap F) \cap G = E \cap (F \cap G)$$

Operations on Sets II

- Distributive Laws

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G) \quad \text{and}$$
$$(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$$

- DeMorgan's Laws

$$(\cup_{i=1}^{n} E_i)^c = \cap_{i=1}^{n} E_i^c \quad \text{and} \quad (\cap_{i=1}^{n} E_i)^c = \cup_{i=1}^{n} E_i^c$$

Mathematically Discussing Probability

- A *probability* is a function that takes events from the sample space Ω and maps them to a value in range $[0, 1]$, that is

$$P : \Omega \longmapsto [0, 1]$$

- A probability satisfies the the following axioms:

Axiom 1	$0 \leq P(E) \leq 1$, for all $E \in \Omega$.
Axiom 2	$P(\Omega) = 1$
Axiom 3	For any sequence of mutually exclusive events $E_1, E_2, \ldots$, we have that

$$P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$$

Example of Probabilities with Dice I

- Consider a fair die with six faces that we will roll once and assess its outcome.
- Our sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Under an assumption that the die is fair, each face should have equal probability of occurring.
  - $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}$ satisfying Axioms 1, 2, & 3.
- By our assumptions and Axiom 3 since each face is mutually exclusive, we can talk about more complex outcomes than just the probability of each face.

Example of Probabilities with Dice II

- For example, we could discuss the probability of odd or even outcomes,

$$\begin{aligned} & P(\{1\}) + P(\{3\}) + P(\{5\}) \\ = \; & P(\{2\}) + P(\{4\}) + P(\{6\}) \\ = \; & \frac{3}{6} = \frac{1}{2} \end{aligned}$$

- Or the probability of an outcome less than 5

$$\begin{aligned} & \text{Pr}(\text{“Rolling a face less than 5”}) \\ = \; & P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) \\ = \; & \frac{4}{6} = \frac{2}{3} \end{aligned}$$

Probability Propositions

- With the Axiom's of probability we can begin to come up with simple relationships that arise out of the axioms.

Proposition (Unproved Probability Propositions)

1. $P(E^c) = 1 - P(E)$
2. If $E \subset F$, then $P(E) \leq P(F)$
3. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

Example of Probability - Another Dice Example I

- These propositions allow us to think of more interesting events that may occur
- For example, we could discuss the probability of NOT rolling a 5.
- We could directly attempt to calculate it, but that would become difficult if we consider sample spaces with large numbers of events.
- Instead, we can appeal to the propositions we just outlined.

Example of Probability - Another Dice Example II

- Now by the first proposition of we have that $P(E^c) = 1 - P(E)$, thus

$$\begin{aligned}
\text{Pr(“Not rolling a 5”)} &= P(\{5\}^c) \\
&= 1 - P(\{5\}) \\
&= 1 - \frac{1}{6} \\
&= \frac{5}{6}
\end{aligned}$$

- We can easily verify this using Axiom 3, where

$$\begin{aligned}
&\text{Pr(“Not rolling a 5”)} \\
&= P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{6\}) = \frac{5}{6}
\end{aligned}$$

From Propositions to Applications

- So why should we can about developing this theoretical understanding of probability
- How does this formalize allow us to talk about real world situations?
- Consider the third proposition where $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
- This proposition can be used to allow us to bound the probability of simultaneous events! (Bonferroni's Inequality)

Bonferroni's Inequality

- *Claim:* $P(E \cap F) \geq P(E) + P(F) - 1.$
- *Proof:* By Proposition (3), $P(E \cup F) = P(E) + P(F) - P(E \cap F)$ and we further can assume that $E \cup F \subseteq \Omega$, thus $P(E \cup F) \leq P(\Omega) = 1$. Therefore

$$P(\Omega) = 1 \geq P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

rearranging this implies that

$$P(E \cap F) \geq P(E) + P(F) - 1$$

While this isn't always useful (notice that we can obtain a negative number for a bound), it shows how theory can start to provide insights in application.

Conditioning on Events

- While probabilities of events are useful, we may often want to talk about events “conditioning” on the fact that we know certain information.
- Conditional probabilities are designed to allow us to calculate a probability given some information.

Defining Conditional Probability

Definition (Conditional Probability)

If $P(F) > 0$, then the conditional probability that E occurs given that F has occurred, denoted $P(E|F)$, is defined

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

- Notice that what happens is that our sample space is now the set $F, (P(F|F) = 1)$
- If the events E and F are disjoint then $P(E|F) = 0$.

Conditional Probability for Multiple Events

- We can extend the definition of conditional probabilities to multiple events

$$
\begin{aligned}
& P(E_1 \cap E_2 \cap E_3 \cap \cdots \cap E_n) \\
= \; & P(E_1) \times P(E_2|E_1) \times P(E_3|E_2 \cap E_1) \times \ldots \\
& \cdots \times P(E_n|E_{n-1} \cap \cdots \cap E_2 \cap E_1)
\end{aligned}
$$

$P(\cdot|F)$ is a probability

Conditional Probabilities also satisfy all of the properties of ordinary probabilities, that is

1) $0 \leq P(E|F) \leq 1$
2) $P(S|F) = 1$
3) If $E_i, i = 1, 2, \ldots$ are mutually exclusive, then

$$P(\cup_{i=1}^{\infty} E_i|F) = \sum_{i=1}^{\infty} P(E_i|F)$$

Therefore after conditioning on certain events we can use all of the probability rules that we have constructed.

Example - Conditional Probability and Cards I

[see Ross, 2014, Example 2a] Joe is 80 percent certain that his missing key is in one of the two pockets of his hanging jacket pocket, being 40 percent certain it is in the left-hand pocket and 40 percent certain it is in the right hand pocket. If the key is not in the left hand pocket, what is the probability that it is in the right pocket.

Example - Conditional Probability and Cards II

Solution

Interested in the "probability the key is in the right pocket, given that it is **not** in the left pocket".

What we know:

- $P(\text{"Left"}) = P(\text{"Right"}) = 0.4$
- Fitting this into the formula for conditional probabilities:

$$P(\text{"Right"} \,|\, \text{"Not Left"}) = \frac{P(\text{"Right" and "Not Left"})}{P(\text{"Not Left"})}$$

- $P(\text{"Not Left"}) = 1 - P(\text{"Not Left"}) = 0.6$
- $P(\text{"Right" and "Not Left"}) = P(\text{"Right"}) = 0.4$

Example - Conditional Probability and Cards III

Thus

$$\begin{aligned} P(\text{“Right”} \,|\, \text{“Not Left”}) &= \frac{P(\text{“Right” and “Not Left”})}{P(\text{“Not Left”})} \\ &= \frac{P(\text{“Right”})}{1 - P(\text{“Left”})} \\ &= \frac{.4}{.6} = \frac{2}{3} \end{aligned}$$

Law of Total Probability

- Consider $E, F \in S$, we can express the event E as follows

$$E = (E \cap F) \cup (E \cap F^c)$$

  where the events $E \cap F$ and $E \cap F^c$ are mutually exclusive (Draw Venn Diagram).

- Thus,

$$\begin{aligned} P(E) &= P(E \cap F) + P(E \cap F^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \end{aligned}$$

- Now, it should be clear that we can extend this to any partition of the set S.

Law of Total Probability

Definition (Law of Total Probability)

Consider the partition $C_1, C_2, \ldots$ of the set S, that is $\cup_{i=1}^{\infty} C_i = S$, additionally consider the event E. Then,

$$\begin{aligned} P(E) &= \sum_{i=1}^{\infty} P(E \cap C_i) \\ &= \sum_{i=1}^{\infty} P(E|C_i)P(C_i) \end{aligned} \tag{1}$$

- The law of total probability will become very important when you begin to investigate Baye's Theorem.

Example 1: Law of Total Probability

Example 1 Suppose you roll a 6-sided die until you get a 6. What is the probability that all numbers rolled are even?

Example 1: Law of Total Probability

Example 1 Suppose you roll a 6-sided die until you get a 6. What is the probability that all numbers rolled are even?

Solution

First, let's identify the sample space and elements of the event of interest.

$$\Omega = \text{Sequences of the form } a_1, a_2, \ldots, a_n \text{ with } a_i \in \{1, 2, 3, 4, 5\}, n \in \mathbb{N}\}$$

For instance,

- 1, 1, 3, 5, 2
- 2, 4
- 4, 2, 2, 4
- 3, 5, 5, 1, 2, 4

The second and third elements in the list contain only evens, and so are in the event of interest.

Example 1: Law of Total Probability

Example 1 Suppose you roll a 6-sided die until you get a 6. What is the probability that all numbers rolled are even?

Solution

The probability of a sequence of rolls containing only 2 and 4 will depend on the length of the sequence. Let the random variable X be the length of the sequence.

We can easily calculate the probability of $X = x$ for a given sequence length x.

$$P(X = x) = \left(\frac{5}{6}\right)^x \frac{1}{6}.$$

Example 1: Law of Total Probability

Example 1 Suppose you roll a 6-sided die until you get a 6. What is the probability that all numbers rolled are even?

Solution

Next, notice that it's easy to calculate the probability of an all even sequence if we know the length of the sequence. Let E be the event that the sequence contains only even numbers.

$$P(E|X = x) = \left(\frac{2}{5}\right)^x.$$

Example 1: Law of Total Probability

Example 1 Suppose you roll a 6-sided die until you get a 6. What is the probability that all numbers rolled are even?

Solution

We can now calculate the desired probability using the law of total probability.

$$
\begin{aligned}
P(E) &= \sum_{x=0}^{\infty} P(E|X=x)P(X=x) \\
&= \sum_{x=0}^{\infty} \left(\frac{2}{5}\right)^{x} \cdot \left(\frac{5}{6}\right)^{x} \frac{1}{6} \\
&= \frac{1}{6} \sum_{x=0}^{\infty} \left(\frac{1}{3}\right)^{x} \\
&= \frac{1}{6} \left(\frac{1}{1-1/3}\right) \\
&= \frac{1}{4}
\end{aligned}
$$

Bayes' Theorem

- Bayes' Theorem allows us to update the probability of an event conditioning on another event happening.

Proposition

Consider a partition $C_1, C_2, \ldots$ of the set S and let E be any set, then for any $i = 1, 2, \ldots$

$$\begin{aligned} P(C_i|E) &= \frac{P(E \cap C_i)}{P(E)} \\ &= \frac{P(E|C_i)P(C_i)}{\sum_{i=1}^{\infty} P(E|C_i)P(C_i)} \end{aligned}$$

where the second line follows from the *law of total probability*

- Used to update beliefs in the presence of new evidence

Example 2: Bayes' Theorem

Example 2 Consider again rolling a die until a six is rolled. What is the probability of a sequence of length x, given that the sequence contains only even numbers?

Solution

Directly apply Bayes' theorem using our previous calculations:

$$\begin{aligned} P(X = x|E) &= \frac{P(E|X = x)P(X = x)}{\sum_{x'=0}^{\infty} P(E|X = x')P(X = x')} \\ &= \frac{(2/5)^x (5/6)^x (1/6)}{\sum_{x'=0}^{\infty} (2/5)^{x'} (5/6)^{x'} (1/6)} \\ &= \frac{\cdot (1/3)^x}{\sum_{x'=0}^{\infty} (1/3)^{x'}} \\ &= \frac{2}{3} \left(\frac{1}{3}\right)^x \end{aligned}$$

Example 3: Bayes' Theorem I

When coded messages are sent, there are sometimes errors in transmission. In particular, Morse code uses "dot" and "dashes", which are known to occur in the proportion 3:4. This means that for an given symbol,

$$P(\text{'dot sent'}) = \frac{3}{7} \quad \text{and} \quad P(\text{'dash sent'}) = \frac{4}{7}$$

Suppose there is interference on the transmission line, and with probability $1/8$ a dot is mistakenly received as a dash, and vice versa. Given that we receive a dot, we are interested in the probability that a dot was sent.

Example 3: Bayes' Theorem II

Solution

Interested in $P(\text{'dot sent'}|\text{'dot received'})$, though now we denote the events as follows

dot sent	$S\circ$
dash sent	$S-$
dot received	$R\circ$
dash sent	$R-$

$$P(R\circ|S\circ) = P(R-|S-) = \frac{7}{8}$$

since with probability $1/8$ a dot is mistaken as a dash. Similarly,

$$P(R\circ|S-) = P(R-|S\circ) = \frac{1}{8}$$

Example 3: Bayes' Theorem III

By Bayes' Theorem, we can write this,

$$\begin{aligned} P(S\circ|R\circ) &= \frac{P(S\circ \cap R\circ)}{P(R\circ)} \\ &= \frac{P(R\circ|S\circ)P(S\circ)}{P(R\circ|S\circ)P(S\circ) + P(R\circ|S-)P(S-)} \\ &= \frac{7/8\times 3/7}{7/8\times 3/7 + 1/8\times 4/7} = \frac{21}{25} \end{aligned}$$

Independence between Events

- Conditional probabilities allow us to talk about how one event occurring changes the probability of another event occurring.
- That is our added knowledge about some event occurring allows us to "update" the probabilities for other events.
- What happens if two events are unrelated though, that is one event is *independent* of another?

Defining Independence

- We now provide a precise definition for what it entails for two events to be independent.

Definition (Independence)

Two events E and F are said to be independent if $P(E \cap F) = P(E)P(F)$. This definition can easily be extended to multiple random variables.

- It is also worth reminding that the concepts of mutually exclusive and independence are different.

Differences between Independent Events and Mutually Exclusive Events

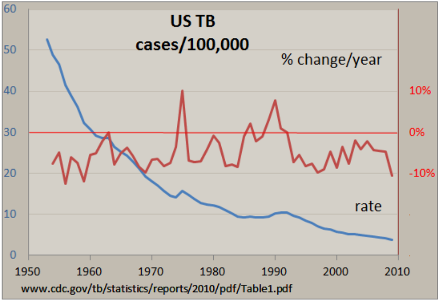
- Recall the definition of mutually exclusive. That is two events are mutually exclusive it $E \cap F$.
- Thus, we can highlight the following consequences and differences based upon definitions and propositions

Mutually Exclusive Events	Independent Events
$P(E \cap F) = 0$	$P(E \cap F) = P(E)P(F)$
$P(E \cup F) = P(E) + P(F)$	$P(E \cup F) = P(E) + P(F) - P(E)P(F)$
$P(E \mid F) = 0$	$P(E \mid F) = P(E)$.

Example Application: Time Series

Time series are an important example of *dependent* data. Modeling time series requires dealing with this dependence in some way, usually by making assumptions about the dependence structure.

Let $x_t, t = 1, \cdots, T$ be values of some quantity observed at time points t. For instance, x_t could be the closing price of a stock over time, daily average temperature, or level of activation in a region of the brain.



Example Application: Time Series

A common time series model is the *autoregressive process* (AR), with the simplest example being the AR(1) process. This process assumes a very specific dependence structure across observations in the time series,

$$P(x_t | x_1, x_2, \ldots, x_{t-1}) = P(x_t | x_{t-1}).$$

This assumption says that the value of x_t depends only on the previous value at time x_{t-1}.
That is, x_t is *conditionally independent* of $x_1, x_2, \cdots, x_{t-2}$ given x_{t-1}.

References

Stephen Abbott. *Understanding analysis*. Springer, 2001.

George Casella and Roger L Berger. *Statistical inference*. Second edition, 2002.

EJ Dudewicz and SN Mishra. Modern mathematical statistics. john wilsey & sons. *Inc., West Sussex*, 1988.

Sheldon Ross. *A first course in probability*. Pearson, ninth edition, 2014.

Mark J Schervish. *Probability and Statistics*. 2014.