STA247

Probability with Computer Applications

Prof. K. H. Wong
Sinan Li

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

First printing, March 2013

Contents

I Contents

Introduction 7
Textbooks 7
Course Information 8
Discussion Board 8

1 Introduction to Probability 11
1.1 Useful Terminology 11

2 Counting 13

3 Conditional Probability 15

4 Discrete Distributions 17
4.1 Random Variable 17
4.2 Cumulative Distribution Function 17
4.2.1 Properties of CDF 17
4.3 Chebyshev's Inequality 18
4.4 Common Discrete Distributions 18
4.4.1 Bernoulli Trials 18

Index 21

Contents

Introduction 7
Textbooks
Course Information
Discussion Board

1 **Introduction to Probability** 11
1.1 Useful Terminology

2 **Counting** 13

3 **Conditional Probability** 15

4 **Discrete Distributions** 17
4.1 Random Variable
4.2 Cumulative Distribution Function
4.3 Chebyshev's Inequality
4.4 Common Discrete Distributions

Index 21

Introduction

This is an introductory course to probability, where our main focus will be developing an understanding of probability and the concept of probability distributions, both for discrete and continuous quantities. This includes developing the intuition for how probabilities 'behave' and the situations in which it is valid to describe randomness using probability, as well as relying on simulations in R to help us visualize these properties.

By the end of the course, students should be able to...

- Describe random quantities in various ways, such as by: their features, density functions, distribution functions, graphs
- Select an appropriate probability model based on their unique properties to quantify the randomness of random quantities
- Compute and interpret the various features of a random quantity: expected value, variance, standard deviation, correlation, covariance, event probabilities (either exactly, through approximations, or simulation)
- Select the most appropriate model to represent randomness and compute probabilities
- Use simulation in R for estimation purposes
- Explain the relationship of transforming a random variable and its effect on its distribution
- Use bivariate distributions to describe the association between two random quantities

Textbooks

We will be referencing these two textbooks regularly:

1. *Probability with Applications and R*, 2nd ed. by *Wagaman and Dobrow* through the library here and with student companion site here.
2. *Modern Mathematical Statistics with Applications*, 3rd ed. by *Devore and Berk* available through the library here.

We will be using R throughout the course to help us understand probability distributions, and to simulate probabilities for quantities that are harder to compute by hand. R Markdown files will be provided for you with necessary starter code. You'll also gain some experience with using LaTeXto produce documents with well-presented math notation.

For new users to R, you may find chapters 3-7 from R for Data Science to be helpful as reference material in data visualization and data manipulation tools in R.

Course Information

All course-related materials can be found on our Quercus page:

- **Lectures** are held in MC 102 for both sections. We begin 10 minutes after the hour, and no recordings will be provided.
- **Weekly materials** such as slides, suggested problems for practice, and reminders of upcoming due dates are posted each week's page on our course home page.
- **Tutorial materials** will be distributed at the beginning of your tutorial.
- **Announcements**: this is the primary channel to distribute important information to everyone! You're expected to check and read announcements regularly to ensure you don't miss any important communication!
- **Assignments** and documents will be distributed and submitted through Crowdmark.

Discussion Board

Throughout the term, beginning September 14, there will be pinned weekly discussions on our discussion board page. They will remain pinned from Wednesday 8 AM to the following Tuesday 8 PM.

- **General Q&A page**: general questions, clarifications, request for additional explanations, share your thoughts/understanding of topics
- **Grouped practice problem discussions**: post your solutions, thoughts, approaches, questions here for that collection of textbook questions.
  - Each discussion thread (i.e. reply to one original comment) is dedicated to ONE question, labeled in bold in the original comment

You earn your discussion board credits by contributing at least five (5) times during this course to any of the discussions during the pinned period (Wednesday to Tuesday) in any of the following ways:

- Posting a question to a problem you tried, with a clear explanation of your process, and if you got stuck: what you tried to do, and where you need help moving forward
- Responding to a question with a thoughtful explanation to help your peer by sharing your own understanding of the problem
- Posting to the general Q&A page with your own question, about a course topic that is still unclear and being specific in describing what you do not understand (and perhaps what you did understand!)
- Providing a detailed and clear response to a question in the general Q&A page

To ensure boards are easy to reference, posting similar content is discouraged and these would be ineligible for earning credits.

- Only contributions during the pinned period will be counted. The discussions will remain open the rest of term for students who come across new problems or would like to continue the discussions.
- You are encouraged to keep the discussion going, but in terms of credit, it will be capped at 5 points.
- While there is no weekly cap on points, the maximum points you can earn in the last two weeks is 2 points, with no more than 2 points per week (i.e. don't wait to the last minute to participate in the class discussion).
- The discussion boards exist to facilitate peer-to-peer collaboration and learning, while also encouraging regular active engagement with course content. The course offers many

opportunities that most students shouldn't find themselves unable to contribute in a unique way.

Why a Discussion Board?

- There are records and studies that have shown the process of explaining and teaching to others is an effective way to learn, consolidate, and retain what has been taught. See here and here.
- It's a space for students to come together to work collaboratively, receive and provide peer support.
- It's also a space to get feedback and guidance from TAs and myself.
- It's a good opportunity to self-assess ('how comfortable am I explaining this to another student?', 'how often do I need to refer back to my notes to explain this concept clearly?') – an important component of good study skills!
- It's valuable information to us! Common questions/misunderstandings that pop up in the weekly discussions can be addressed during our weekly lecture meetings.

Discussion Board Rubric

Points	1 point	0.5 points	0 points
Quality of contribution	Student has made a substantial and unique contribution with detailed explanations and/or clearly outlined process of the approach to a problem. Student was involved in follow-up discussions and worked collaboratively with their peers to develop a better understanding of the concepts involved.	Student has made a contribution to the discussion that is dismissive, lacking in detail, or not completely unique. Unable to further the discussion in a way that fosters a collaborative learning environment. e.g. responses such as 'you just need to integrate this and solve for it' or 'I got the same answer doing... (reiterates OPs process)'	Student has not contributed to the weekly discussion topic thread, or whose posts are off-topic/irrelevant/do not contribute to the thread or is not unique to what has already been discussed in the thread. e.g. 'I got the same answer', 'How did you get that number?'

Tutorial

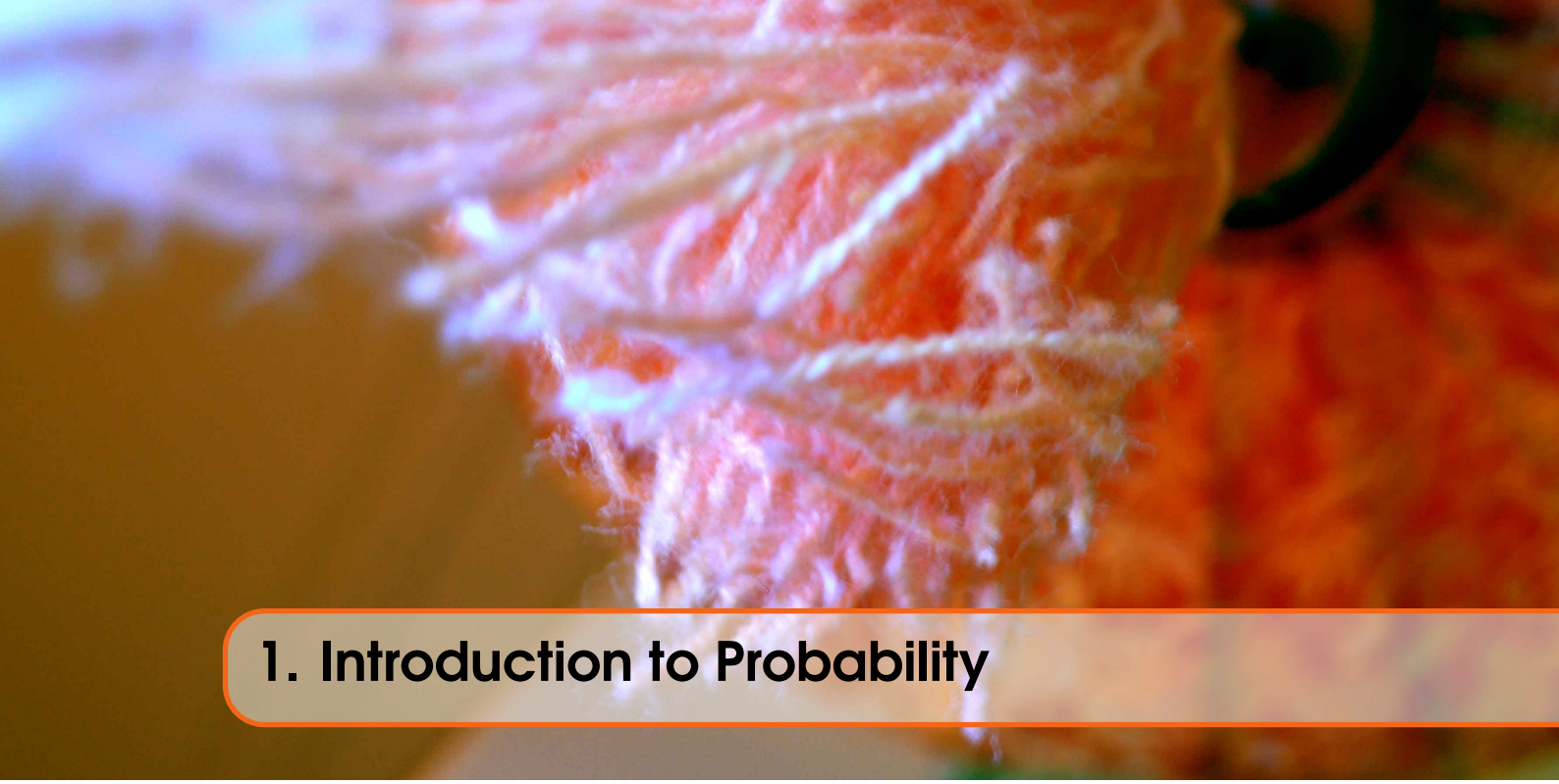
- A mix of R labs and collaborative pair work
- R labs: These labs have guided exercises to practice the R tools covered in class or learn new R skills. Labs will be TA-guided.
- R labs require individual .rmd and knitted document submissions at the end of tutorial (Note that the labs are guided and meant to be completed within the tutorial time)
- Pair work: In your tutorial section, you will with a partner of your choice work on more challenging but guided problems together. Discussion and sharing your ideas is a great way to learn from one another, and consider different approaches to problem solving! TAs will be there to support and help answer clarification questions.

Habits for Success

- Attend and participate in lecture. Try to work along with the problems presented. Ask questions and interact with your peers during the open work periods!
- Make sure you focus on *understanding* the concepts and how they relate and build upon each other. This one is hard!
- Regularly attempt as many of the suggested problems as you can and work towards being able to work on the problems closed book. Use this to gauge your familiarity with the material – if it takes you an hour to work through two problems, then it's a good indication to seek out advice and support from the teaching team. The earlier, the better!
- Drop by during the office hours or post on discussion board if you get stuck. Work through practice problems with classmates. Take turns *explaining* your thinking and problem solving process.
- **Create a schedule and stick with it**. This course covers many topics to ensure you have a good foundation for latter courses. Many topics build on top of each other so falling behind can quickly snowball and make it difficult to catch up.

Some Suggestions from Previous Students

1. Will the final be cumulative?
   Yes.
2. I find that I *really* struggle with keeping my mind focused on the question I'm presenting working on, and not think about how a similar question was solved previously. I thought I was making some progress, but then I realized this wasn't the case at all after today when I spent 30 minutes trying to recall the solution I wrote for A2 (I don't remember even reading the test question..I just caught a few phrases and began regurgitating my assignment solution incorrectly).
   I realize the most helpful thing I can do for myself now is to practice regularly, but I can't say I've been doing that well. For the last week, I've been attempting the textbook questions (open book)in preparation for the test. For the final, I plan to solve questions independently, which was something I didn't do until last minute. What else can I do to improve my problem-solving/critical thinking skills?
   Definitely do textbook / slide questions closed book. The only thing you should be looking at is the sample aid sheet. Otherwise you will never know if you actually understand the material.
3. You need to know when you don't know something. Meaning that when you see a question, you need to recognize if you actually know how to solve it or not. And if you don't know how to solve it you need to skip it and go to the next question. I'm sure you know that spending 30 min on any question is not an efficient use of the time.
4. Go to office hours. There is only like 2 other people that I've ever seen at office hours.
5. What does data have to look like for an exponential distribution to describe it?
   Why do we use different distributions? What's the importance of justifying a decision to use one distribution over another? (the textbook often tells you the distribution so image all of the questions posed to you without that information. Would you still be confident?) Answering these questions thoroughly will make you a lot better at figuring out how to approach a question when you see it.
   Understanding the conceptual side in depth is the most important part of studying I would think. Computational skills come with practice, eventually you get it. But usually the problem with stats or courses like this is interpreting the question - is figuring out how to even approach the question in the first place. Knowing how to solve all of the normal distribution questions that confront you is great, but that won't help you if you can't even recognize when and when not to use it.



1. Introduction to Probability

1.1 Useful Terminology

Definition 1.1.1 — Random Experiment. A process that allows us to gather data or observations. Experiment can be repeated multiple times under the same conditions. The set of possible outcomes of the experiment are known, but the outcome of a specific experiment is not known.

■ **Example 1.1** Below are some examples of random experiments.

- Rolling a die and observing the top-facing number
- Rolling a pair of dice and observing the sum of top-facing numbers
- A patient being administered a painkiller and observing the amount of time in minutes before relief is felt

■

Definition 1.1.2 — Sample Space.

2. Counting

3. Conditional Probability

4. Discrete Distributions

4.1 Random Variable

4.2 Cumulative Distribution Function

The probability behaviour of a random variable can be represented in many ways, such as with the probability mass function. Another representation is with the *cumulative distribution function*.

Definition 4.2.1 — Cumulative Distribution Function. The *cumulative distribution function* (CDF) $F(x)$ of a discrete random variable with probability mass function $P(x)$ or $f(x)$ is a function that returns the cumulative (total) probability up to and including $X = x$.

$$F(b) = P(X \leq b) = \sum_{x \in \{x \leq b\}} P(x)$$

The domain of the CDF is always over the set of real numbers! As such, CDFs are often represented as a piecewise function.

■ **Example 4.1** Find the cumulative distribution function for PMF below:

x	0	1	2	3
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

$$F(x) = \begin{cases} \frac{1}{6} \\ \frac{2}{3} \\ \frac{29}{30} \\ 1 \end{cases}$$

■

4.2.1 Properties of CDF

CDF of a Discrete Random Variable

For a discrete random variable X with CDF $F(X)$:

1. The graph of the CDF will be a **non-decreasing step-function**. That is for $a < b$, $F(a) \leq F(b)$.
2. The graph of the CDF is **right continuous**. That is, $\lim_{x \to c^+} F(x) = F(c)$.
3. $\lim_{x \to \infty} F(x) = 1$
4. $\lim_{x \to -\infty} F(x) = -1$

Add Stuff Here (Week 5 page 17 - 20)

4.3 Chebyshev's Inequality

Theorem 4.3.1 — Chebyshev's Inequality. Let X be a random variable with mean (expected value) μ and finite variance σ^2. Then for any positive k,

$$P(|x - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

Proof. By Markov's Inequality: for non negative x, $P(x \geq a) \leq \frac{E(x)}{a}$, $a > 0$.
$P(|X - \mu| < k\sigma) = P((x-\mu)^2 < k^2\sigma^2)$ since RV's are non-nagative
FINISH THE PROOF ■

■ **Example 4.2** Based on past data, the average daily number of tech support requests at a local call centre is 115 with a standard deviation of 10 calls.

a) What can be said about the fraction of days on which the number of calls received is between 100 and 130?
Dist info: missing
We are given: $\mu = 115$, $r = 10$
Let C be the random number of daily calls.

$$\begin{aligned} P(100 \leq C \leq 130) &= P(-15 \leq C - 115 \leq 15) \\ &= P(-15 \leq C - \mu \leq 15) \\ &= P(|C - \mu| \leq 15) \\ &= P(|C - \mu| < 16) \\ &= P(|C - \mu| < 1.6\sigma) \\ &\geq 1 - \frac{1}{1.6} = 0.6094 \end{aligned}$$

$\therefore$ At least 60.94% of the time they will have between 100 to 130 calls a day.

b) What number of calls can they expect to receive at least 90% of the time?

■

4.4 Common Discrete Distributions

4.4.1 Bernoulli Trials

Definition 4.4.1 — Bernoulli Trials. A *Bernoulli trial* is a random experiment consisting of exactly one trial involving two possible outcomes, often called a *success* or a *failure*. Let X be the outcome of a Bernoulli trial where FINISH on PAGE 24

Often, we are interested in modeling the number of successes among multiple trials instead of the results of a single trial:

Definition 4.4.2 — Binomial Distribution. A *Binomial experiment* consists of n independent and identical Bernoulli trials. The probability of success, p, is fixed for each trial.

Let X be the random variable representing the number of successes among the n trials. Then X can be modeled by the binomial distribution with parameters n and p, denoted as $X \sim \text{Bin}(n, p)$. The binomial distribution has probability mass function:

$$O(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

If $X \sim \text{Bin}(n, p)$, we can show that $E(X) = np$ and $V(X) = np(1-p)$

Index

B

Bernoulli Trials 18
Binomial Distribution 18

C

Chebyshev's Inequality 18
Cumulative Distribution Function 17

R

Random Experiment 11

S

Sample Space 11