MATH 697

Sahir Rai Bhatnagar

2017-11-21

Contents

Syllabus

General Information

- Instructor(s): Sahir Bhatnagar and Dr. Alexandra M. Schmidt
- Email: sahir.bhatnagar@mail.mcgill.ca,
- Website: http://sahirbhatnagar.com/MATH697/
- Lectures: Tuesdays 9am - 12pm
- Office: TBD
- Office Hours: By appointment only
- Prerequisite(s): Calculus and Algebra
- Texts: *Modern Mathematical Statistics with Applications*, 2nd Edition by Jay L. Devore and Kenneth N. Berk

Course Description

The main learning outcomes of this course are to get a broad idea about some frequently used probability models and to learn basic results and techniques in probability theory and statistical inference. Most of the materials for the course will be drawn from the first seven chapters of the textbook. The book does not, however, contain all the materials we intend to cover in this course. Some extra notes will therefore be given on those topics not in the text book. We will also introduce computational methods in statistics with the statistical software program `R`.

Grade Distribution

Assignments	10%
Quizzes	40%
Final Exam	50%

Target Syllabus

Overview and Descriptive Statistics (Weeks 1-4)

- 1.1 Populations and Samples

Probability (Weeks 1-4)

- 2.1 Sample Spaces and Events
- 2.2 Axioms, Interpretations, and Properties of Probability
- 2.3 Counting Techniques
- 2.4 Conditional Probability
- 2.5 Independence

Discrete Random Variables and Probability Distributions (Weeks 1-4)

- 3.1 Random Variables
- 3.2 Probability Distributions for Discrete Random Variables
- 3.3 Expected Values of Discrete Random Variables
- 3.4 Moments and Moment Generating Functions
- 3.5 The Bernoulli/Binomial Probability Distribution
- 3.6 The Geometric/Negative Binomial Probability Distribution
- 3.7 The Poisson Probability Distribution

Continuous Random Variables and Probability Distributions (Weeks 5-8)

- 4.1 Probability Density Functions and Cumulative Distribution Functions
- 4.2 Expected Values and Moment Generating Functions
- 4.3 The Uniform Distribution
- 4.4 The Exponential Distribution
- 4.5 The Gamma Distribution
- 4.6 The Normal Distribution
- 4.7 One-Dimensional Change of Variable (Discrete and Continuous)

Joint Probability Distributions (Weeks 5-8)

- 5.1 Jointly Distributed Random Variables
- 5.2 Expected Values, Covariance, and Correlation
- 5.3 Conditional Distributions
- 5.4 Multidimensional Change of Variable (Discrete and Continuous)

Sampling Distributions and Limits (Weeks 5-8)

- 6.1 Sampling Distributions
- 6.2 Convergence in Probability, Weak Law of Large Numbers
- 6.3 Convergence with Probability 1, Strong Law of Large Numbers
- 6.4 Convergence in Distribution, Central Limit Theorem

Statistical Inference (Weeks 9-12)

- 7.1 Inference Using a Probability Model
- 7.2 Statistical Models
- 7.3 Data Collection (Finite Populations, Simple Random Sampling, Histograms)
- 7.4 Basic Inferences (Descriptive, Plots, Types of Inferences)

Likelihood Inference (Weeks 9-12)

- 8.1 The Likelihood Function, Sufficient Statistics
- 8.2 Maximum Likelihood Estimation
- 8.3 Inferences Based on the MLE (Standard Errors, Bias, Consistency, Confidence Intervals, Hypotheses and Test Procedures, P-values, Inferences for the Variance)
- 8.4 Distribution-Free Methods (Method of Moments, Bootstrapping)

Regression and Correlation (Weeks 9-12)

- 9.1 The Simple Linear and Logistic Regression Models
- 9.2 Estimating Model Parameters
- 9.3 Inferences About the Regression Coefficient
- 9.4 Inferences Concerning Prediction of Future Y Values
- 9.5 Correlation
- 9.6 Model Checking (χ^2 Goodness of Fit Test, Cross-Validation)
- 9.7 Multiple Regression Analysis
- 9.8 Regression with Matrices

Prerequisites

Install R and RStudio

All examples in this book are run in an R environment. You also need a recent version of RStudio, which is a software application that facilitates how you interact with `R`. It is developed by data enthusiasts who consider statistics to be more than just simulations, formulas and proofs. RStudio emphasizes the following:

1. **Version control**: Why I should use version control especially for the solo data analyst.
2. **Reproducible research**: seamless integration with RMarkdown for creating dynamic documents and presentations
3. **Creating R Packages**: seamless integration with the devtools package for creating software that implements your statistical method or analysis.

R Packages

The following packages will be called upon at some point, so please install them before getting started with the tutorials. Enter the following command in `R`:

```
install.packages(c("pacman", "knitr", "data.table", "rmarkdown", "tidyverse",
    "boot", "Hmisc"))
```

Introduction to R

Try out the interactive tutorial: `http://swirlstats.com/`

Background Reading

The greatest thing about `R` is that there are so many people out there willing to help you. `R` users are constantly writing tutorials and creating packages to make your analysis tasks easier. Here is a very targeted list that I suggest reading prior to starting the tutorials

1. Writing Functions
2. `for` loops
3. `apply` vs. `for`

Slides

1. Discrete Random Variables and Probability Distributions (part I)
2. Discrete Random Variables and Probability Distributions (part II)
3. Continuous Random Variables and Probability Distributions
4. Normal Distribution and Expectations of Continuos Random Variables
5. Transformations of a Random Variable and Discrete Joint Distributions
6. Joint, Marginal, Conditional Continuous Distributions
7. Multidimensional Change of Variables, Conditional Expectation, Variance, Hierarchical Distributions
8. Sampling Distributions and Limits, Convergence in Probability
9. Convergence in Distribution and Central Limit Theorem

Assignments

1. A1 (due September 26, 2017)
2. A2 (due October 12, 2017)
3. A3 (due October 26, 2017)
4. A4 (due November 9, 2017)

Quiz

1. Quiz 1 (October 3, 2017)
2. Quiz 2 (November 7, 2017)

R Code

0.1 Central Limit Theorem in Action

1. clt-master.Rmd
2. clt-template.Rmd

Part I

Part I

Chapter 1

Overview and Descriptive Statistics[1]

Statistical concepts and methods are not only useful but indeed often indispensable in understanding the world around us. They provide ways of gaining new insights into the behavior of many phenomena that you will encounter in your chosen field of specialization.

The discipline of statistics teaches us how to make intelligent judgments and informed decisions in the presence of *uncertainty and variation*.

Without uncertainty or variation, there would be little need for statistical methods or statisticians.

If the yield of a crop were the same in every field, if all individuals reacted the same way to a drug, if everyone gave the same response to an opinion survey, and so on, then a **single observation would reveal all desired information**.

1.1 Populations and Samples

We are constantly exposed to collections of facts, or data, both in our professional capacities and in everyday activities. The discipline of statistics provides methods for organizing and summarizing data and for drawing conclusions based on information contained in the data.

An investigation will typically focus on a well-defined collection of objects constituting a population of interest: - In one study, the population might consist of all gelatin capsules of a particular type produced during a specified period. - Another investigation might involve the population consisting of all individuals who received a B.S. in mathematics during the most recent academic year.

When desired information is available for all objects in the population, we have what is called a census. Constraints on time, money, and other scarce resources usually make a census impractical or infeasible. Instead, a subset of the population, **a sample**, is selected in some prescribed manner. Thus we might obtain a sample of pills from a particular production run as a basis for investigating whether pills are conforming to manufacturing specifications, or we might select a sample of last year's graduates to obtain feedback about the quality of the curriculum.

1.1.1 Variable

We are usually interested only in certain characteristics of the objects in a population: the amount of vitamin C in the pill, the gender of a mathematics graduate, the age at which the individual graduated, and so on.

[1] Devore and Berk.

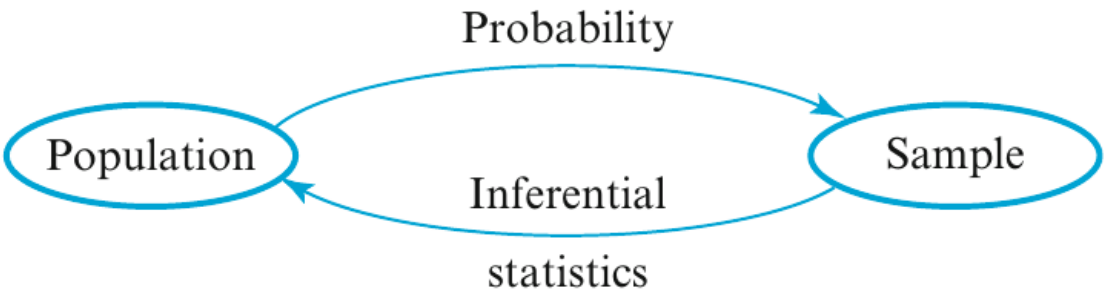

Figure 1.1

A variable is any characteristic whose value may change from one object to another in the population. Can be categorical (male/female) or numerical (temperature).

data type	description
univariate	consists of observations on a single variable
bivariate	observations are made on each of two variables
multivariate	more than two variables

1.1.2 Branches of Statistics

- **Descriptive Statistics**: summarize and describe important features of the data. Can be graphs (histograms, boxplots, and scatter plots), or numeric summaries (mean, standard deviations, and correlation coefficients)
- **Inferential Statistics**: Techniques for generalizing from a sample to a population. Having obtained a sample from a population, an investigator would frequently like to use sample information to draw some type of conclusion (make an inference of some sort) about the population. That is, the sample is a means to an end rather than an end in itself.

The focus of this couse is Inferential statistics. But to get there we need to understand the basic concepts of probability

The relationship between the two disciplines can be summarized by saying that probability reasons from the population to the sample (**deductive reasoning**), whereas inferential statistics reasons from the sample to the population (**inductive reasoning**).

Before we can understand what a particular sample can tell us about the population, we should first understand the uncertainty associated with taking a sample from a given population. This is why we study probability before statistics.

Example 1.1 (Use of manual lap belts in cars equipped with automatic shoulder belt systems)**.** *Probability*: assume that 50% of all drivers in a certain metropolitan area regularly use their lap belt → an assumption about the population. We might ask - How likely is it that a sample of 100 such drivers will include at least 70 who regularly use their lap belt? - How many of the drivers in a sample of size 100 can we expect to regularly use their lap belt?

Inference: a sample of 100 drivers of such cars revealed that 65 regularly use their lap belt. We might ask - Does this provide substantial evidence for concluding that more than 50% of all such drivers in this area

regularly use their lap belt We are attempting to use sample information to answer a question about the structure of the entire population from which the sample was selected.

1.2 Pictorial and Tabular Methods in Descriptive Statistics

1.3 Measures of Location

1.4 Measures of Variability

Chapter 2

Probability

Introduction[1]

The random variation associated with *measurement* procedures in a scientific analysis requires a framework in which the **uncertainty** and **variability** that are inherent in the procedure can be handled. The key goal of Probability and Statistical modelling is to establish a mathematical framework within which *random* variation (due, for example, to experimental error or natural variation) can be quantified so that *systematic* variation (arising due to potentially important biological differences) can be studied.

Broadly, the *Scientific Process* involves several different stages:

$$\begin{array}{ccc} \text{THEORETICAL MODELLING} & \rightarrow & \text{MATHEMATICAL/PROBABILISTIC MODELLING} \\ \downarrow & & \\ \text{PREDICTION} & & \\ \downarrow & & \\ \text{EXPERIMENTATION/OBSERVATION} & & \\ \downarrow & & \\ \text{VALIDATION} & & \end{array}$$

Mathematical/Probabilistic modelling facilitates PREDICTION; *Statistical Analysis* provides the means of validation of predicted behaviour.

To explain the variation in observed data, we need to introduce the concept of a *probability distribution*. Essentially we need to be able to model, or specify, or compute the *chance* of observing the data that we collect or expect to collect. This will then allow us to assess how likely the data were to occur by chance alone, that is, how *surprising* the observed data are in light of an assumed theoretical model.

For example, consider two nucleotide sequences of the same length that we wish to assess for similarity:

Example 2.1 (Two nucleotide sequences)**.**

Sequence 1 $ATAGTAGATACGCACCGAGGA$

Sequence 2 $ATCTTAGATAGGCACTGAGGA$

How can we assess sequence similarity formally ? The number of discordant positions is 4, but how informative is that summary measure ? Perhaps we need to assess the chance, for example, that a point

[1] Reproduced with permission from http://www.math.mcgill.ca/dstephens/

mutation

$$A \to C$$

occurs (as in the discordant position 3) in unit evolutionary time. Perhaps the chance of observing a subsequence

$$ATCTTA$$

rather than

$$ATAGTA$$

(in positions 1-6) is important.

- Is the hidden (or *latent*) structure in the sequence, corresponding to whether the sequence originates from a coding region or otherwise, important ?
- Can we even infer the hidden structure in light of the data we have observed ?

These questions can only really be answered when we have an understanding of randomness and variation. The framework that we will use to pose and answer such questions formally is given to us by *probability theory*.

Probability: A Measure of Uncertainty[2]

Often in life we are confronted by our own ignorance. Whether we are pondering tonight's traffic jam, tomorrow's weather, next week's stock prices, an upcoming election, or where we left our hat, often we do not know an outcome with certainty. Instead, we are forced to guess, to estimate, to hedge our bets.

> Probability is the science of uncertainty.

It provides precise mathematical rules for understanding and analyzing our own ignorance. It does not tell us tomorrow's weather or next week's stock prices; rather, it gives us a **framework for working with our limited knowledge** and for **making sensible decisions based on what we do and do not know**.

To say there is a 40% chance of rain tomorrow is not to know tomorrow's weather. Rather, it is to **know what we do not know** about tomorrow's weather. In this course, we will develop a more precise understanding of what it means to say there is a 40% chance of rain tomorrow. We will learn how to work with ideas of randomness, probability, expected value, prediction, estimation, etc., in ways that are sensible and mathematically clear.

2.1 Sample Spaces and Events

2.1.1 Sample Spaces

Definition 2.1 (Sample Space)**.** The sample space Ω is the set of possible outcomes of an experiment. Points ω in Ω are called sample outcomes, realizations, or elements.

Example 2.2 (Coin tossing)**.** $\Omega = \{H, T\}$

Example 2.3 (Dice)**.** $\Omega = \{1, 2, 3, 4, 5, 6\}$

Example 2.4 (Proportions)**.** $\Omega = \{x : 0 \leq x \leq 1\}$

[2] http://www.utstat.toronto.edu/mikevans/jeffrosenthal/book.pdf

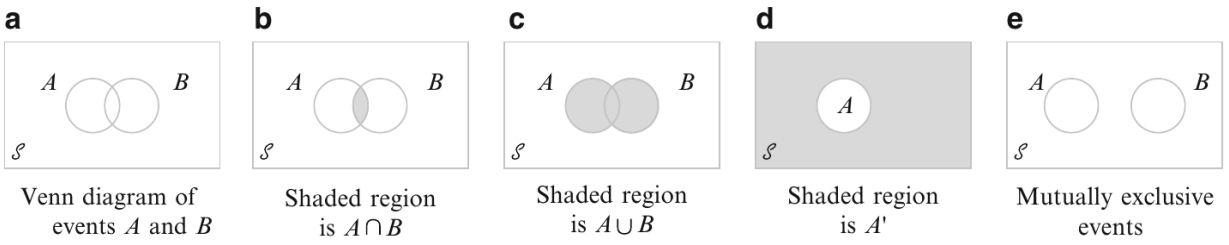

Figure 2.1

Example 2.5 (Time measurement)**.** $\Omega = \{x : x > 0\} = \mathbb{R}^+$

Example 2.6 (Temperature measurement)**.** $\Omega = \{x : a \leq x \leq b\} \subseteq \mathbb{R}$

Example 2.7 (Biological Sequence Analysis)**.** The experiment may involve the observation of a nucleotide or protein sequence, so that the sample space Ω may comprise all sequences (of bases/amino acids) up to a given length, and a sample outcome would be a particular observed sequence.

There are two basic types of experiment: - Counting - Measurement

We shall see that these two types lead to two distinct ways of specifying probability distributions.

The collection of sample outcomes is a **set** (a collection of items) written as

$$s \in \Omega$$

if s *is a member of the set* Ω.

2.1.2 Events

Definition 2.2 (Event)**.** An event E is a subset of the sample space Ω ($E \subseteq \Omega$). Events are usually denoted by upper case letters near the beginning of the alphabet, like A, B, C. An event which consists of only one outcome is called a simple (or elementary event); otherwise it is a compound event.

The sets Ω and E can be either be written as a list of items, for example,

$$E = \{s_1, s_2, ..., s_n, ...\}$$

which may a finite or infinite list, or can only be represented by a continuum of outcomes, for example

$$E = \{x : 0.6 < x \leq 2.3\}$$

Events are manipulated using **set theory** notation; if A and B are two events, $A, B \subseteq \Omega$, then

- $A \cup B$ is the set of outcomes that belong to A **or** to B, or to both,
- $A \cap B$ is the set of outcomes that belong to both A **and** to B.
- A^c (complement of A) is the set of outcomes **not** in A
- $A \backslash B = A \cap B^c$

The empty event will be denoted by $\varnothing$. Two events A and B are mutually exclusive if $A \cap B = \varnothing$, i.e., the collection of sample outcomes have no element in common.

2.2 Axioms, Interpretations, and Properties of Probability

Definition 2.3 (Axioms (basic properties) of Probability)**.** To ensure that the probability assignments will be consistent with our intuitive notions of probability, all assign- ments should satisfy the following axioms (basic properties) of probability

- **AXIOM 1:** For any event A,
$$P(A) \geq 0$$
- **AXIOM 2:**
$$P(\Omega) = 1$$
- **AXIOM 3:** If $A_1, A_2, \ldots$ is an infinite collection of disjoint events, then
$$P\left(A_1 \cup A_2 \cup \cdots\right) = \sum_{i=1}^{\infty} P(A_i)$$

Proposition 2.1. $P(\varnothing) = 0$ where $\varnothing$ is the null event. This in turn implies that the property contained in Axiom 3 is valid for a finite collection of events.

Proposition 2.2. For any event A,

$$P(A) = 1 - P(A^c)$$

Proposition 2.3. For any event A,

$$P(A) \leq 1$$

Proposition 2.4. For any events A and B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

2.3 Counting Techniques

When the various outcomes of an experiment are equally likely (the same probability is assigned to each simple event), the task of computing probabilities reduces to counting. In particular, if N is the number of outcomes in a sample space and $N(A)$ is the number of outcomes contained in an event A, then

$$P(A) = \frac{N(A)}{N}$$

Proposition 2.5 (Product rule for ordered pairs)**.** If the first element or object of an ordered pair can be selected in n_1 ways, and for each of these n_1 ways the second element of the pair can be selected in n_2 ways, then the number of pairs is $n_1 \cdot n_2$.

2.3.1 Permutations

Definition 2.4 (Permutation)**.** Any ordered sequence of k objects taken from a set of n distinct objects is called a permutation of size k of the objects. The number of permutations of size k that can be constructed from the n objects is denoted by P_k^n:

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

2.3.2 Combinations

Definition 2.5 (Combination)**.** Given a set of n distinct objects, any unordered subset of size k of the objects is called a combination. The number of combinations n of size k that can be formed from n distinct objects will be denoted by

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

2.4 Conditional Probability

Conditional probability is the means by which probabilities are updated in the light of new information. We examine how the information *an event B has occurred* affects the probability assigned to A.

Example 2.8 (Flipping Coins)**.** We flip three different fair coins, and

$$\Omega = HHH, HHT, HTH, HTT, THH, THT, TTH, TTT$$

with $P(s) = 1/8$ for each $s \in \Omega$. What is the probability that the first coin comes up heads?

$$P(\text{first coin heads}) = P(HHH, HHT, HTH, HTT) = 4/8 = 1/2$$

But suppose now that an informant tells us that exactly two of the three coins came up heads. Now what is the probability that the first coin was heads? if we know that exactly two of the coins were heads, then we know that the outcome was one of HHT, HTH, THH.

Because those three outcomes should (in this case) still all be equally likely, and because only the first two correspond to the first coin being heads, we conclude the following: If we know that exactly two of the three coins are heads, then the probability that the first coin is heads is 2/3.

More precisely, we have computed a conditional probability. That is, we have determined that, conditional on knowing that exactly two coins came up heads, the conditional probability of the first coin being heads is 2/3. We write this in mathematical notation as

$$P(\text{first coin heads}|\text{two coins heads}) = 2/3.$$

Here the vertical bar | stands for *conditional on* or *given that*.

Example 2.9 (Assembly Lines)**.** Complex components are assembled in a plant that uses two different assembly lines, A and A^c . Line A uses older equipment than A^c, so it is somewhat slower and less reliable. B are the defective components and B^c are the nondefective.

	Condition	
Line	B	B^c
A	2	6
A^c	1	9

The sales manager randomly selects 1 of these 18 components for a demonstration

$$P(\text{line A component was selected}) = P(A) = \frac{N(A)}{N} = \frac{8}{18} = 0.444$$

However, if the chosen component turns out to be defective, then the event B has occurred, so the component

must have been 1 of the 3 in the B column of the table. Since these 3 components are equally likely among themselves after B has occurred

$$P(\text{line A component was selected}|\text{Defective}) = \frac{2}{3} = \frac{2/18}{3/18} = \frac{P(A \cap B)}{P(B)}$$

In Example 2.9, the conditional probability is expressed as a ratio of **unconditional probabilities**. The numerator is the probability of the intersection of the two events, whereas the denominator is the probability of the conditioning event B. Given that B has occurred, the relevant sample space is no longer Ω but consists of just outcomes in B; A has occurred if and only if *one of the outcomes in the intersection* occurred, so the conditional probability of A given B is proportional to $P(A \cap B)$. The proportionality constant $1/P(B)$ is used to ensure that the probability $P(B|B)$ of the new sample space B equals 1.

Definition 2.6 (Conditional Probability)**.** Given two events A and B, with $P(B) > 0$, the conditional probability of A given B is equal to

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

In example 2.8, let
- $A = HHH, HHT, HTH, HTT$ be the event that the first coin is heads
- $B = HHT, HTH, THH$ be the event that exactly two coins were heads

It follows that

$$A \cap B = HHT, HTH$$

Therefore

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(HHT, HTH)}{P(HHT, HTH, THH)} = \frac{2/8}{3/8} = \frac{2}{3}$$

Example 2.10 (Balanced die)**.** Suppose a balanced die is tossed in the next room. We are told that a number less than 4 was observed. What is the probability the number was either 1 or 2?

Example 2.11 (Two Balanced Dice v1)**.** Toss two balanced dice. Let $A = \{\text{sum of 5}\}$ and $B = \{\text{first die is} \leq 2\}$. Find $P(A|B)$

Example 2.12 (Two Balanced Dice v2)**.** Two balanced dice are tossed. What is the probability that the first die gives a number less than three, given that the sum is odd?

Example 2.13 (Unbalanced Die)**.** Toss an unbalanced die with probs $P(1) = .1$, $P(2) = .1$, $P(3) = .3$, $P(4) = .2$, $P(5) = .1$, $P(6) = .2$. Let $A = \geq 5$ and $B = \geq 2$. Find $P(A|B)$.

Example 2.14 (Two Balanced Coins)**.** Two balanced coins were tossed, and it is known that at least one was a head. What is the probability that both were heads?

Example 2.15 (Two Cards)**.** Two cards are drawn without replacement from a standard deck. Find the probability that

1) the second is an ace, given that the first is not an ace.

2) the second is an ace.

3) the first was an ace, given that the second is an ace.

Example 2.16 (Numbers in a Hat)**.** The numbers 1 to 5 are written on five slips of paper and placed in a hat. Two slips are drawn at random without replacement. What is the probability that the first number is 3, given a sum of seven?

Example 2.17 (One Card)**.** A card is selected at random (i.e. every card has the same probability of being chosen) from a deck of 52. What is the probability it is a red card or a face card?

Definition 2.6 immediately leads to the *multiplication formula*

Definition 2.7 (Multiplicative Rule)**.**

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

and

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

This allows us to compute the joint probability of A and B when we are given the probability of B and the conditional probability of A given B, and vice versa.

Example 2.18 (Fish in a Tank)**.** A tank has three red fish and two blue fish. Two fish are chosen at random and without replacement. What is the probability of getting

1) red fish first and then a blue fish?

2) both fish red?

3) one red fish and one blue fish?

2.4.1 Law of Total Probability

Recall that events $A_1, A_2, \ldots, A_k$ are mutually exclusive if no two have any common outcomes. The events are exhaustive if one A_i must occure, so that

$$A_1 \cup A_2 \cup \cdots \cup A_k = \Omega$$

Theorem 2.1 (Law of Total Probability)**.** *Let $A_1, A_2, \ldots, A_k$ be mutually exclusive and exhaustive events. Then for any other event B,*

$$P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

Proof: Because the A_i's are mutually exclusive and exhaustive, if B occurs it must be in conjunction with exactly one of the A_i's. That is, $B = (A_1 \text{ and } B)$ or $\ldots$ or $(A_k \text{ and } B)$ which is equal to $(A_1 \cap B) \cup \cdots \cup (A_k \cap B)$, where the events $(A_i \cap B)$ are mutually exclusive.

Thus we have

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

Example 2.19 (Long Hair)**.** Suppose a class contains 60% girls and 40% boys. Suppose that 30% of the girls have long hair, and 20% of the boys have long hair. A student is chosen uniformly at random from the class. What is the probability that the chosen student will have long hair?

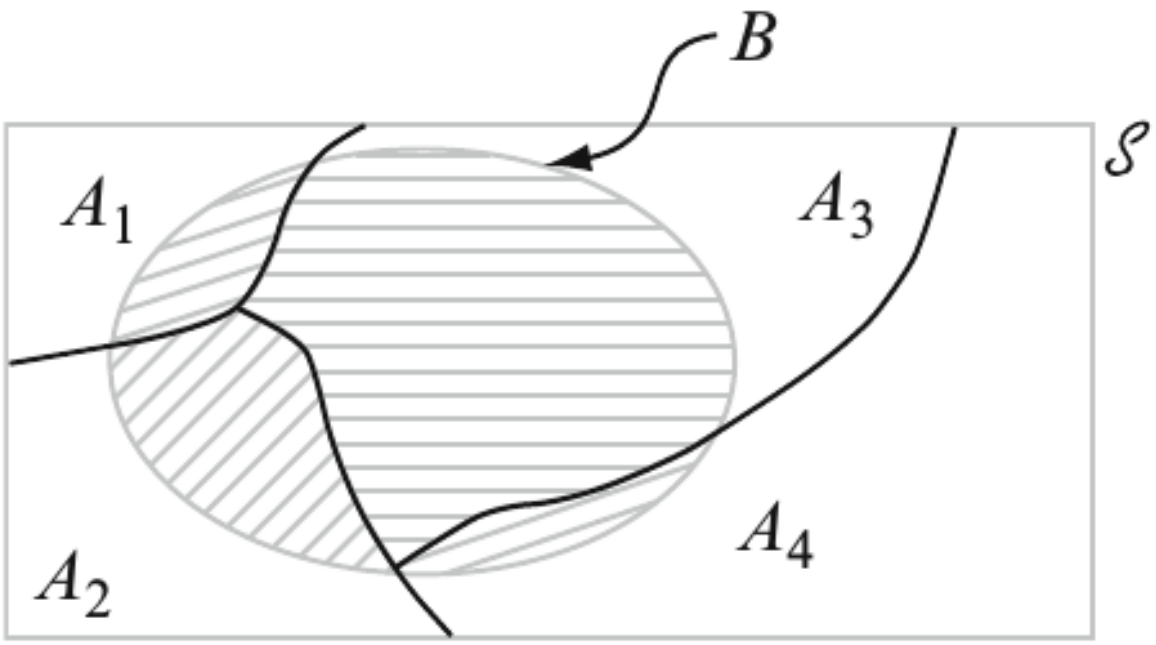

Figure 2.2

2.4.2 Bayes' Rule

Theorem 2.2 (Bayes' Rule)**.** *Let $A_1, \ldots, A_k$ be a collection of mutually exclusive and exhaustive events with $P(A_i) > 0$ for $i = 1, \ldots, k$. Then for any other event B, for which $P(B) > 0$, we have*

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}, \quad j = 1, \ldots, k$$

The transition from the second to the third expression in Theorem 2.2 rests on using the multiplication rule in the numerator and the law of total probability in the denominator.

Example 2.20 (Urns)**.** Suppose urn #1 has 3 red and 2 blue balls, and urn #2 has 4 red and 7 blue balls. Suppose one of the two urns is selected with probability 1/2 each, and then one of the balls within that urn is picked uniformly at random.

1) What is the probability that urn #2 is selected at the first stage (event A) and a blue ball is selected at the second stage (event B)?

2) Compute the probability that a blue ball is obtained.

3) Now suppose we are given the information that the ball picked is blue. What is probability that we had selected urn #2?

Example 2.21 (Large Bridges)**.** There are three Canadian firms which build large bridges, firm 1, firm 2, and firm 3. 20% of Canadian large bridges have been built by firm 1, 30% by firm 2, and the rest by firm 3. 5% of the bridges built by firm 1 have collapsed, while 10% of those by firm 2 have collapsed, and 30% by firm 3 have collapsed.

1) What is the probability that a bridge collapses?

2) Suppose it is reported in tomorrow's newspaper that a large bridge has collapsed. What is the probability it was built by firm 1?

2.5 Independence

If we flip a fair coin twice, then the probability of two heads is $1/2 \times 1/2$. We multiply the probabilities because we regard the two tosses as independent. The formal definition of independence is as follows:

Definition 2.8 (Indepent Events)**.** Two events A and B are independent if

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

Now, because $P(A|B) = P(A \cap B)/P(B)$, we see that A and B are independent if and only if $P(A|B) = P(A)$ or $P(B|A) = P(B)$, provided that $P(A) > 0$ and $P(B) > 0$. Definition 2.8 has the advantage that it remains valid even if $P(B) = 0$ or $P(A) = 0$, respectively. Intuitively, events A and B are independent if neither one has any impact on the probability of the other.

Example 2.22 (Toss a fair coin 10 times)**.** Toss a fair coin 10 times. Let $A =$ at least one head. Let T_j be the event that tails occurs on the j^{th} toss. Find $P(A)$

Example 2.23 (Unbalance Die Revisited)**.** In Example 2.13, if A is the event that the die was 5, and B is the event that the coin was tails, then calculate $P(A), P(B)$ and $P(A \cap B)$

Chapter 3

Discrete Random Variables and Probability Distributions

Introduction

In Chapter 2, we discussed the probability model as the central object of study in the theory of probability. This required defining a probability measure P on a class of subsets of the sample space Ω. For example, for an experiment with possible sample outcomes denoted by the *sample space* Ω, an *event* E was defined as any collection of sample outcomes, that is, any subset of the set Ω.

$$\begin{array}{lll} \text{EXPERIMENT} & \longrightarrow \text{SAMPLE OUTCOMES} & \longrightarrow \text{EVENTS} \quad \longrightarrow \text{PROBABILITIES} \\ & \Omega = \{s_1, s_2, ...\} & \longrightarrow E \subseteq S \quad \longrightarrow P(E) \end{array}$$

In this framework, it is necessary to consider each experiment with its associated sample space separately - the nature of sample space Ω is typically different for different experiments.

Example 3.1 (Rainy days)**.** Count the number of days in February which have zero precipitation.

$$\Omega = \{0, 1, 2, \ldots, 28\}$$

Let $E_i = i$ days have zero precipitation. $E_0, \ldots, E_{28}$ partition Ω.

Example 3.2 (Footbal Match)**.** Count the number of goals in a football match.

$$\Omega = \{0, 1, 2, 3, \ldots\}$$

Let $E_i = i$ goals in the match. $E_0, E_1, E_2, \ldots$ partition Ω

In both of these examples, we need a formula to specify each

$$P(E_i) = p_i$$

Example 3.3 (Operating Temperature)**.** Measure the operating temperature of an experimental process.

$$\Omega = \{x : x > T_{min}\}$$

Here it is difficult to express

$$P(\text{“Measurement is } x)$$

but possible to think about

$$P(\text{“Measurement is } \leq x) = F(x)$$

and now we seek a formula for $F(x)$ which is a simpler way of presenting a particular probability assignment.

This chapter is concerned with the definitions of random variables, distribution functions $F(x)$, probability/density functions $f(x)$, and the development of the concepts necessary for carrying out calculations for a probability model using these entities (Evans and Rosenthal, 2004).

The concept of a random variable allows us to pass from the experimental outcomes themselves to a numerical function of the outcomes. There are two fundamentally different types of random variables (Devore and Berk, 2011):

i) discrete random variables

ii) continuous random variables

In this chapter, we examine the basic properties and discuss the most important examples of **discrete** variables. Chapter 4 focuses on continuous random variables.

3.1 Random Variables

A general notation useful for all such examples can be obtained by considering a sample space that is **equivalent** to Ω for a general experiment, but whose form is more familiar.

Definition 3.1 (Random Variable)**.** A random variable X on Ω is a function from the sample space Ω to the set $\mathbb{R}$ of all real numbers denoted by

$$X : \Omega \to \mathbb{R}$$

Let R_X denote the range of X.

X is called a *discrete* random variable if R_X is a countable set.

Random variables are customarily denoted by uppercase letters, such as X and Y, lowercase letters to represent some particular value of the corresponding random variable. The notation $X(s) = x$ means that x is the value associated with the outcome s by the rv X.

Example 3.4 (Coin Toss)**.** Suppose a coin is tossed three times. Let X be the number of heads observed. The sample space is

$$\Omega = \left\{ \underbrace{HHT}_{3}, \underbrace{HHT}_{2}, \underbrace{HTH}_{2}, \underbrace{HTT}_{1}, \underbrace{THH}_{2}, \underbrace{THT}_{1}, \underbrace{TTH}_{1}, \underbrace{TTT}_{0} \right\}$$

That is, we have $X(HHH) = 3$, $X(HHT) = 2$, $X(HTH) = 2$, and so on. Hence $R_X = \{0, 1, 2, 3\}$

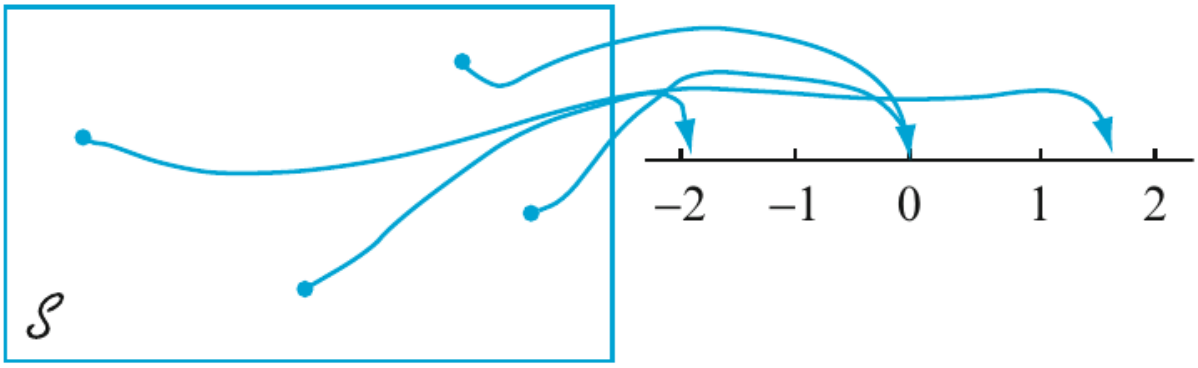

Figure 3.1 A random variable

Figure 3.1

Example 3.5 (A Very Simple Random Variable)**.** Let the random variable $X : \{\text{rain, snow, clear}\} \to \mathbb{R}$ by $X(rain) = 3$, $X(snow) = 6$, and $X(clear) = -2.7$.

We now present several further examples. The point is, we can define random variables any way we like, as long as they are functions from the sample space to $\mathbb{R}$.

Example 3.6 (A Very Simple Random Variable 2)**.** For the case $\Omega = \{\text{rain, snow, clear}\}$, we might define a second random variable Y by saying that $Y = 0$ if it rains, $Y = -1/2$ if it snows, and $Y = 7/8$ if it is clear. That is $Y(rain) = 0$, $Y(snow) = 1/2$, and $Y(rain) = 7/8$.

Example 3.7 (A Very Simple Random Variable 3)**.** If the sample space corresponds to flipping three different coins, then we could let X be the total number of heads showing, let Y be the total number of tails showing, let $Z = 0$ if there is exactly one head, and otherwise $Z = 17$.

Example 3.8 (Constants as Random Variables)**.** As a special case, every constant value c is also a random variable, by saying that $c(s) = c$ for all $s \in \Omega$. Thus, 5 is a random variable, as is 3 or -21.6.

Example 3.9 (Indicator Functions)**.** If A is any event, then we can define the indicator function of A, written I_A, to be the random variable

$$I_A(s) = \begin{cases} 1 & s \in A \\ 0 & s \notin A \end{cases}$$

Suppose X is a random variable. We know that different states s occur with different probabilities. It follows that $X(s)$ also takes different values with different probabilities. These probabilities are called the **distribution** of X; we consider them next.

3.2 Probability Distributions for Discrete Random Variables

Because random variables are defined to be functions of the outcome s, and because the outcome s is assumed to be random (i.e., to take on different values with different probabilities), it follows that the value of a random variable will itself be random (as the name implies).

Specifically, if X is a random variable, then what is the probability that X will equal some particular value x? Well, $X = x$ precisely when the outcome s is chosen such that $X(s) = x$.

3.3 Expected Values of Discrete Random Variables

3.4 Moments and Moment Generating Functions

3.5 The Binomial Probability Distribution

3.6 The Poisson Probability Distribution

Chapter 4

Continuous Variables and Probability Distributions

Introduction

Appendix A

Vectorization, *apply and for loops

This section will cover the basics of vectorizations, the `*apply` family of functions and `for` loops.

A.1 Vectorization

Almost everything in `R` is a vector. A scalar is really a vector of length 1 and a `data.frame` is a collection of vectors. An nice feature of `R` is its vectorized capabilities. Vectorization indicates that a function operates on a whole vector of values at the same time and not just on a single value[1]. If you have have ever taken a basic linear algebra course, this concept will be familiar to you.
Take for example two vectors:

$$\begin{bmatrix}1\\2\\3\end{bmatrix} + \begin{bmatrix}1\\2\\3\end{bmatrix} = \begin{bmatrix}2\\4\\6\end{bmatrix}$$

The corresponding `R` code is given by:

```
a <- c(1, 2, 3)
b <- c(1, 2, 3)
a + b
```

```
## [1] 2 4 6
```

Many of the `base` functions in `R` are already vectorized. Here are some common examples:

```
# generate a sequence of numbers from 1 to 10
(a <- 1:10)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
# sum the numbers from 1 to 10
sum(a)
```

```
## [1] 55
```

[1] http://www.dummies.com/how-to/content/how-to-vectorize-your-functions-in-r.html

```r
# calculate sums of each column
colSums(iris[, -5])
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        876.5        458.6        563.7        179.9
```

Exercise: What happens when you sum two vectors of different lengths?

A.2 Family of *apply functions

- `apply`, `lapply` and `sapply` are some of the most commonly used class of functions in `R`
- `*apply` functions are not necessarily faster than loops, but can be easier to read (and vice cersa)
- `apply` is used when you need to perform an operation on every row or column of a matrix or data.frame
- `lapply` and `sapply` differ in the format of the output. The former returns a list while the ladder returns a vector
- There are other `*apply` functions such as `tapply`, `vapply` and `mapply` with similar functionality and purpose

A.2.1 Loops vs. Apply

```r
# Getting the row means of two columns Generate data
N <- 10000
x1 <- runif(N)
x2 <- runif(N)
d <- as.data.frame(cbind(x1, x2))
head(d)
```

```
##           x1        x2
## 1 0.54792654 0.1450859
## 2 0.42577875 0.7253068
## 3 0.01186565 0.6229836
## 4 0.54296831 0.5309603
## 5 0.09945893 0.9147921
## 6 0.98512496 0.2707034
```

```r
# Loop: create a vector to store the results in
rowMeanFor <- vector("double", N)

for (i in seq_len(N)) {
    rowMeanFor[[i]] <- mean(c(d[i, 1], d[i, 2]))
}

# Apply:
rowMeanApply <- apply(d, 1, mean)

# are the results equal
all.equal(rowMeanFor, rowMeanApply)
```

```
## [1] TRUE
```

A.2.2 Descriptive Statistics using *apply

```
data(women)
# data structure
str(women)
```

```
## 'data.frame':    15 obs. of  2 variables:
##  $ height: num  58 59 60 61 62 63 64 65 66 67 ...
##  $ weight: num  115 117 120 123 126 129 132 135 139 142 ...
```

```
# calculate the mean for each column
apply(women, 2, mean)
```

```
##   height   weight
##  65.0000 136.7333
```

```
# apply 'fivenum' function to each column
vapply(women, fivenum, c(Min. = 0, `1st Qu.` = 0, Median = 0, `3rd Qu.` = 0,
    Max. = 0))
```

```
##         height weight
## Min.      58.0  115.0
## 1st Qu.   61.5  124.5
## Median    65.0  135.0
## 3rd Qu.   68.5  148.0
## Max.      72.0  164.0
```

A.2.3 Creating new columns using sapply

You can apply a *user defined function* to columns or the entire data frame:

```
# the ouput of sapply is a vector the 's' in sapply stands for 'simplified'
# apply
mtcars$gear2 <- sapply(mtcars$gear, function(i) if (i == 4) "alot" else "some")

head(mtcars)[, c("gear", "gear2")]
```

```
##                   gear gear2
## Mazda RX4            4  alot
## Mazda RX4 Wag        4  alot
## Datsun 710           4  alot
## Hornet 4 Drive       3  some
## Hornet Sportabout    3  some
## Valiant              3  some
```

A.2.4 Applying functions to subsets using tapply

```
# Fisher's famous dataset
data(iris)
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 ...
```

```
# mean sepal length by species
tapply(iris$Sepal.Length, iris$Species, mean)
```

```
##     setosa versicolor  virginica
##      5.006      5.936      6.588
```

A.2.5 Nested for loops using `mapply`

`mapply` is my favorite `base R` function and here are some reasons why:

- Using `mapply` is equivalent to writing nested `for` loops except that it is 100% more human readable and less prone to errors
- It is an effective way of conducting simulations because it iterates of many arguments

Let's say you want to generate random samples from a normal distribution with varying means and standard deviations. Of course the brute force way would be to write out the command once, copy paste as many times as you want, and then manually change the arguments for `mean` and `sd` in the `rnorm` function as so:

```
v1 <- rnorm(100, mean = 5, sd = 1)
v2 <- rnorm(100, mean = 10, sd = 5)
v3 <- rnorm(100, mean = -3, sd = 10)
```

This isn't too bad for three vectors. But what if you want to generate many more combinations of means and sds ? Furthermore, how can you keep track of the parameters you used? Now lets consider the `mapply` function:

```
means <- c(5, 10, -3)
sds <- c(1, 5, 10)

# MoreArgs is a list of arguments that dont change
randomNormals <- mapply(rnorm, mean = means, sd = sds, MoreArgs = list(n = 100))

head(randomNormals)
```

```
##          [,1]      [,2]       [,3]
## [1,] 5.937124  9.303791  0.6359922
## [2,] 5.326156 20.523454  8.8685494
## [3,] 7.057136  8.670827 12.7212667
## [4,] 4.174939  8.324977 -4.0635612
## [5,] 4.885643 15.276630 -4.2899465
## [6,] 5.923214 13.078001 19.5706231
```

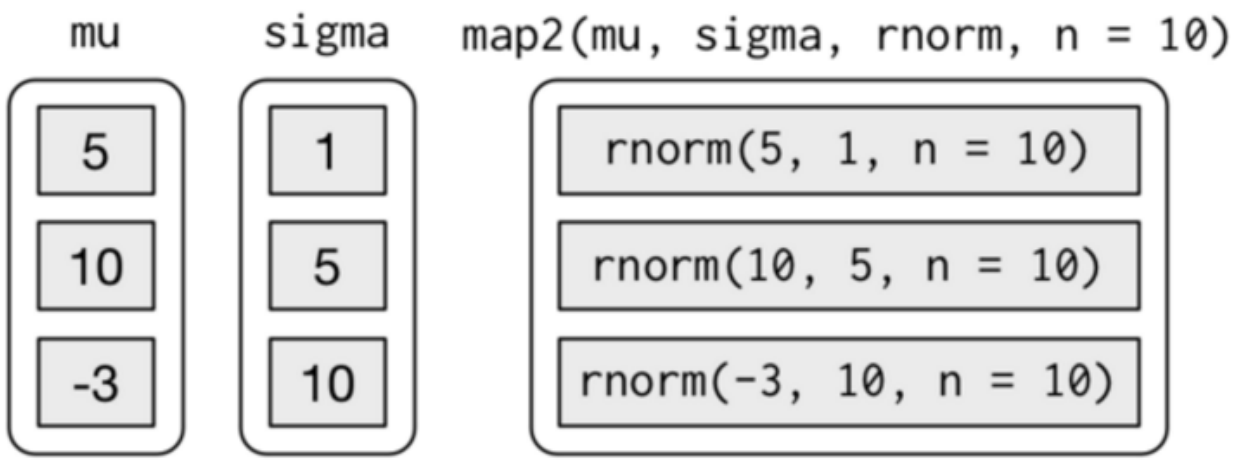

Figure A.1

The following diagram (from r4ds) describes exactly what is going on in the above function call to `mapply`:

Advantages:

1. Result is automatically stored in a matrix
2. The parameters are also saved in `R` objects so that they can be easily manipulated and/or recovered

Consider a more complex scenario where you want to consider many possible combinations of means and sds. We take advantage of the `expand.grid` function to create a `data.frame` of simulation parameters:

```
simParams <- expand.grid(means = 1:10, sds = 1:10)

randomNormals <- mapply(rnorm, mean = simParams$means, sd = simParams$sds, MoreArgs = list(n = 100))

dim(randomNormals)
```

```
## [1] 100 100
```

A.3 Creating dynamic documents with `mapply`

`mapply` together with the `rmarkdown` package (Allaire et al., 2017) can be very useful to create dynamic documents for exploratory analysis. We illustrate this using the Motor Trend Car Road Tests data which comes pre-loaded in `R`.

> The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Copy the code below in a file called `mapplyRmarkdown.Rmd` :

Copy the code below in a file called `boxplotTemplate` :

Appendix B

Appendix B

Bibliography

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., and Chang, W. (2017). *rmarkdown: Dynamic Documents for R*. R package version 1.7.

Devore, J. L. and Berk, K. N. (2011). *Modern Mathematical Statistics with Applications*. Springer Science & Business Media.

Evans, M. J. and Rosenthal, J. S. (2004). *Probability and statistics: The science of uncertainty*. Macmillan.