Contents

1 Characterizing Data

1.1 Introduction

Solving Statistical Problem

The first two steps of solving a statistical problem are:

1. **The Problem.** Identify the problem or objective. What *questions* would you like to answer? What *hypotheses* do you have?
2. **The Data.** Decide how to take a *sample* from your *population* and what data to collect on that sample. This step is informed by **the problem**.

This class is not about these first two steps. It is about what comes next. Today, we are concerned with characterizing data, presumably generated by someone diligently following the two first steps above. Today's data example will be data collected on you!

1.2 Graphical Methods

1.2.1 R: Loading Data

R: Loading Data

```
# read the sample file of student data into R
> d <- read.csv(file="example.csv", header=T)
> print(d)
   First.Name Mid.Name  Last.Name Major Clsfn.Year College
1    Nicholas    James   Abdallah  MATH          3       S
2       David  Michael     Bloyer  MATH          3       S
3       Kevin  Michael       Born  MATH          4       S
...
```

- `read.csv`, `read.table`, and related functions are used to read data from text files
- Data files are formatted like big matrices. The first row is often a **header** giving the name of the columns. The remaining rows contain the data measurements for each element in your sample. The columns are separated by spaces (`read.table`), tabs (`read.table` with argument `sep="\t"`), or commas (`read.csv`).
- **csv** files: data files where the columns are separated by commas

1.2.2 R: Working with Data Frames

Once data has been read into a data frame, you will want to access its columns and rows. If the data file contained a header, then the columns will be named according to these headers. To get all the middle names in the class, for example, you would access the `Mid.Name` column

```
 [1] James   Michael Michael Dolan   L       Kirk            Glen
[10] Sik     Chih    Bing            Anita   Michael Jean    Robert  Matthew
[19] Alan    Daniel  James   James   Hailu   Mehmet  Ann     Raymond Russell
[28]
21 Levels:  Alan Anita Ann Bing Chih Daniel Dolan Glen Hailu James ... Sik
```

You can ignore the bit about `Levels` until you take a class in statistics that covers ANOVA. We see there are 28 middle names, some of them blank or, in one case, just an initial.

To access a particular record, for example, the person with only the middle initial, we notice that we want to extract the data for the 5th individual. Here, we treat the data frame like a matrix and access the 5th row, all columns:

```
> d[5,]
  First.Name Mid.Name Last.Name Major Clsfn.Year College
5      Renee        L    Dunkin  MATH          3       S
```

We could get fancy and access the `Major` of all students with middle name `Michael`:

```
> d$Major[d$Mid.Name=="Michael"]
[1] MATH  MATH  CPR E
13 Levels: AER E CH E COM S CPR E ECON ENSCS FIN I E L ST MATH M E ... STAT
```

to see that there are 3 students with middle name "Michael", 2 math majors and 1 computer engineering major.

1.2.3 R: Calling Functions

R: Calling Functions

Equivalent commands in `R`:

```
d <- read.csv(file="example.csv", header=T)
d <- read.csv(header=T, file="example.csv")
d <- read.csv(file="example.csv")
```

- You will use *functions* in `R` to carry out calculations and operations.
- **Arguments**. Functions take multiple arguments that are named. For example, `read.csv` takes an argument called `file` and another called `header`. Arguments can be listed in any order, as long as you refer to them by their name. Many arguments have default values, so you do not need to provide them if the default works for you. For example, `header` defaults to `T`, aka `TRUE`.
- **Return Value**. Functions return objects. For example, `read.csv` returns what is called a *data frame*, which is basically an internal representation of the file. You can capture the return value of a function and store it in a *variable* by using the *assignment* operator `<-`. Above, `d` is a variable that holds the data frame returned by `read.csv`.

1.2.4 R: Getting Help

R: Getting Help

- If you need more help to use a function, type a question mark plus the name of the command, for example `?read.csv`.
- If you don't know the name of the command, you can search for terms in the `R` help. For example, if you remember "csv", but not the name of the command, you might search `help.search("csv")`. The output indicates which functions help files match the search string.

```
Help files with alias or concept or title matching 'csv' using
regular expression matching:



read.table(utils)       Data Input
write.table(utils)      Data Output



Type 'help(FOO, package = PKG)' to inspect entry
'FOO(PKG) TITLE'.
```

Now, you would type `?read.table` and it will tell you about `R` commands that work with csv files.

Characterizing Measurements

How you choose to characterize your data is up to you. It depends a lot the **the problem**.

There are two phases of data characterization

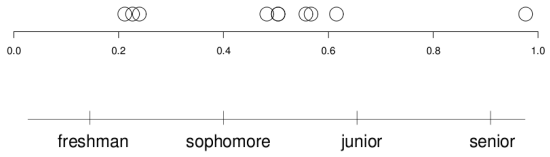
1. **Exploratory**. Use your eyes and graphical methods to *explore* your data. You aren't necessarily focused on a question, but it helps to keep in mind Step 1 **the problem** to guide your exploration, and it helps to use your imagination. This field is among the most artistic within statistics or science in general.
2. **Inference-Directed**. Now you focus your attention on answering a particular question. You summarize the data in the way that most directly answers the question. You also consider what theory you know because you'll need to understand the uncertainty of your characterization, i.e. your chances of being wrong with your answer to the question. Not all characterizations are equally suited for answering questions, and you need some theoretical knowledge to guide you choice.

Example: Choosing Characterizations

Suppose you want to determine whether students taking this class tend to be upperclassmen. Notice the column `Clsfn.Year` in the `example.csv` data set.

- A *graphical characterization* can be obtained by plotting a histogram (`hist` in `R`).
- There are multiple *numerical characterizations*, including:
  - **average**: the bigger the average `Clsfn.year`, the more upperclassmen or the more senior upperclassmen in the class
  - **proportion of upperclassmen**: the number of `Clsfn.year`≥ 3 divided by the class size directly addresses the question, and is probably the best summary of the data for the question at hand

1.2.5 The Histogram



The measurements in your sample sit on a segment of the real line. Or if they are discrete, such as freshman, sophomore, etc., you can line them up at equally spaced intervals along a line segment. The point is, your data can be viewed as sitting in a line.

Making a Histogram

With this in mind, here are the steps to drawing a histogram:

- Divide the line into (equally-sized) bins.
- Count how many measurements land in each bin.
- Plot a bar graph, with each bar sits above one bin. The height of the bars is either:
  - **Fequency Histogram**. $H =$ number in the bin
  - **Relative Fequency Histogram**.

$$H = \frac{\text{number in bin}}{\text{sample size} \times W}$$

where W is the width of the bin.

Area Under Rectangle

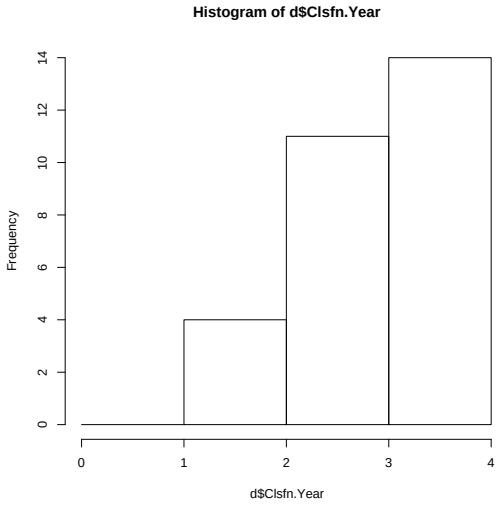
Notice that the area of a rectangle is

$$\text{area of rectangle} = H \times W = \frac{\text{number in bin}}{\text{sample size}}$$

which is just the proportion of the sample that falls in this bin. Equivalently, we can think of this area as the probability that a randomly selected sample measurement falls within the limits of this bin.

1.2.6 R: Histograms

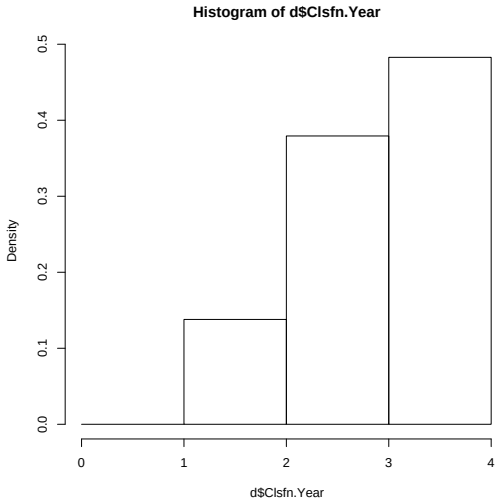
Example: R Frequency Histogram




```
> hist(x=d$Clsfn.Year, breaks=c(0,1,2,3,4))
```

The R function `hist` plots histograms. By default, it plots the **frequency histogram**.

Example: R Relative Frequency Histogram



```
> hist(x=d$Clsfn.Year, breaks=c(0,1,2,3,4), freq=F)
```

With argument `freq=F` set to false, it plots the **relative frequency histogram**.

R: `hist` Function

- The `x` argument are the data for which you would like a histogram.
- The `breaks` argument is optional. It defines the boundaries of the bins. Usually good defaults are chosen. However, for this particular example, I did not like the defaults, and chose my own breaks.
- There are many other arguments (see `?hist` for more information) that you can use to tweak the way the histogram is plotted.

1.3 Numerical

1.3.1 Introduction

Numerical Summaries

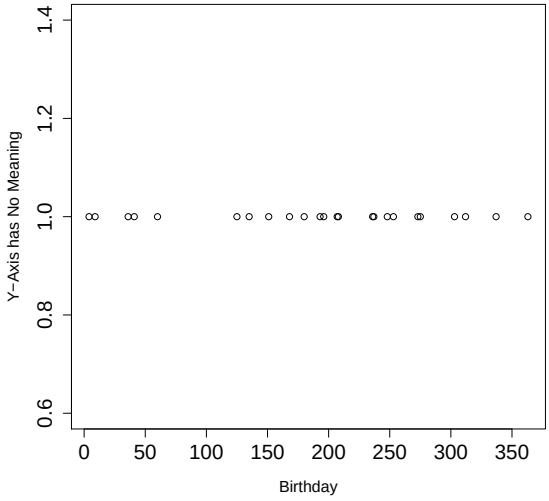
We have seen how the **histogram** can be used to summarize a dataset graphically. Now we will talk about numerical summaries.

A *statistic* is a function that maps the data of a sample to a real number. Numerical summaries of datasets are therefore statistics.

There are infinitely many functions that could be used to compute statistics from samples. You choose the one(s) that are most informative for answering the question at hand.

Example: characterizing birthdays

Suppose you wanted to know if there are non-random patterns in the birthdays of this class. In other words, you want to know if birthdays are like random draws from the 365 days in a year. What kind of statistics would be useful for addressing this question? Let's consider some ways that birthdays can be non-random to guide us in our search for reasonable statistics. An exploratory analysis (see figure) of the data might suggest some ideas of patterns.



- Are birthdays unusually clumped? Some statistics that inform on the degree of "clumping" in the data are:

  standard deviation (we'll see it shortly), minimum days between two consecutive birthdays, average number of days between consecutive birthdays, maximum proportion of birthdays on the same day, in the same week, or in the same month

- Are birthdays unusually spread out, for example one every 20th day? Some statistics that inform on the regularity of the data are:

  standard deviation of the days between consecutive birthdays, minimum days between consecutive birthdays, maximum days between consecutive birthdays, maximum number or proportion of birthdays happening on the same day, same week, same month

There is some overlap in the statistics informative on both questions, but it is interesting to note that this is not true in all cases. It takes a fair amount of creativity and thinking to come up with good statistics. And coming up with a statistic is only half the battle. For inference, you'll also need to understand the properties of the statistic.

1.3.2 Sample Mean

Sample Mean $\bar{x}$

Definition: *sample mean*

Suppose your sample consists of data $y_1, y_2, \ldots, y_n$. Then, the sample mean is defined as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

1.3.3 Sample Variance

Sample Variance s^2

Definition: *sample variance*

With the same sample notation as above, the sample variance is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

where $\bar{y}$ is the sample mean.

Notice, we divide by $n-1$ rather than n. This is done because it turns out s^2 is a better estimate of the population variance (that we define shortly), but you won't see the proof until Stat 342.

Sometimes, we refer to the standard variance of data set $y_1, y_2, \ldots, y_n$ as s_y^2 to be precise about dataset beging referred to.

Sample Standard Deviation

Definition: *sample standard deviation*

The sample standard deviation is

$$s = \sqrt{s^2}$$

The nice thing about the standard deviation is that its units match those of the original data. For example, sample variance of birthday data is measured in days^2, but the sample standard deviation is measured in days.

1.3.4 Population Parameters

Population Parameters

The sample mean and variance can be computed on any sample dataset. If we could sample the *entire* population (which we almost never can), then we could compute the corresponding population statistics. Because these numbers are special, we give them a different name and refer them to as *population parameters* (not statistics). Because we can almost never actually measure all units in the population, these numbers are considered *unknowable*. We can estimate, guess, hypothesize what they are, but we can rarely know what they are for sure.

Definition: *population mean* μ

> The population mean is μ. If the population is finite of some large (but untenable) size N, then $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$ for measurements y_i on all individuals.

Definition: *population variance* σ^2

> The population variance is σ^2. If the population is finite,
>
> $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$$

The population standard deviation is $\sigma = \sqrt{\sigma^2}$.

Empirical Rule

Rule: *Empirical Rule*

> For a broad range of conditions, you can expect the following proportions of your data to fall in the following ranges

Percent of Observations	Range
68%	$(\mu - \sigma, \mu + \sigma)$
95%	$(\mu - 2\sigma, \mu + 2\sigma)$
nearly 100%	$(\mu - 3\sigma, \mu + 3\sigma)$

This rule is not a provable theorem because it is only true when the data come from the canonical "bell-shaped curve" (i.e. the normal distribution). However, because so many different types of data come from a distribution that is fairly "bell-like", the rule applies roughly to many kinds of data.

1.3.5 R: Computing Summary Statistics

R: `mean()`, `sd`, `var`

Suppose you have read data into an `R` object you've called `data`. And suppose there is a column of data in that object called `data$V1`. Then, you can easily compute numerical summaries of the data.

Function	What it does
`mean()`	Computes the sample mean of the data provided in its first argument
`sd()`	Computes the sample standard deviation of the data provided in its first argument
`var()`	Computes the sample variance of the data provided in its first argument

Part I
Probability

2 Introduction

Randomness

Almost everything we observe has some degree of randomness. Just because life is random does not mean we do not understand many aspects of life very well. We have two types of knowledge about randomness:

- We know all the possible outcomes (i.e. we know the *sample space*).
  - We knew that either Obama or McCain would win on November 4.
  - We know a quarter, when flipped, will either come up tails or heads.
- We can guess the probability that a particular outcome will occur.
  - People were prognosticating the elections throughout last year and quantifying the probability of an Obama win in may ways (probability, odds, or even price at the Iowa Electronic Markets).
  - Most of us would simply agree that a fair coin has equal probability of landing heads or tails.

Calculating Probabilities

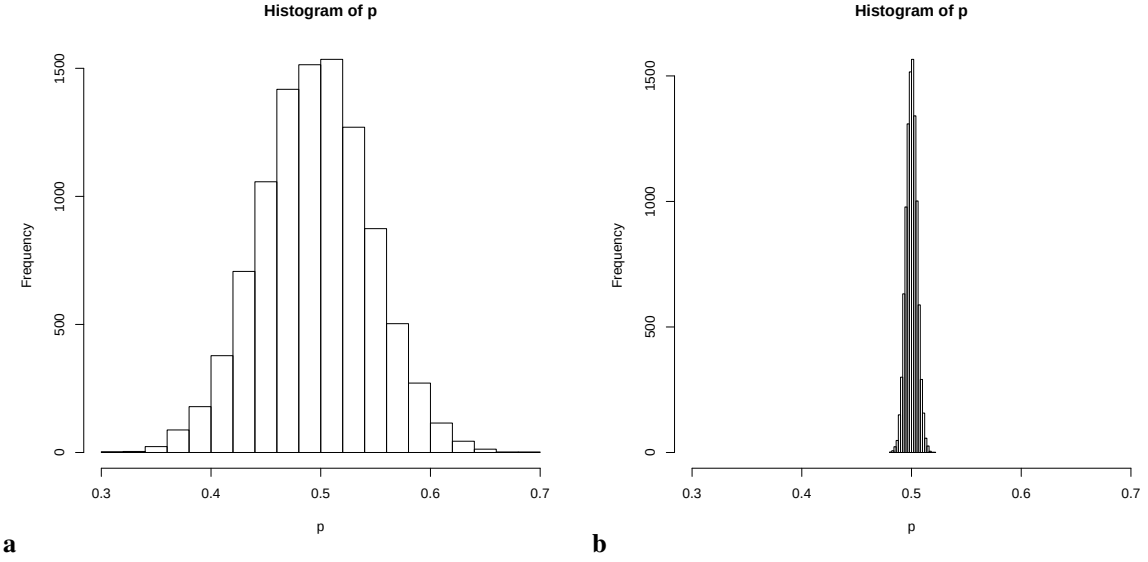
Fair coins aside, how do we come up with numeric probabilities? It can be a tough process. In this class, we will discuss situations where the probabilities can be computed exactly (in other words, we deal with very controlled situations). In general, there are three ways to get at probabilities:

- **Knowledge and Common Sense.** A lot of probabilities can be computed simply by knowing. For example, a team of physicist and biologists could probably give you some pretty convincing evidence that there is a 50:50 probability of getting a head or a tail upon flipping a coin, but you already had that common sense. You could also probably tell me my chances of drawing a red diamond from a deck of cards without too much trouble. For the latter calculation, you are using a counting method that we will discuss later.
- **Experience and Experimentation.** If you study something for a long time, you begin to see patterns. You can estimate the future probability of a particular outcome by computing the proportion of time the same outcome occurred in the past. For example, if you are anxious about taking tests, perhaps it is because you've had a few bad experiences in the past, in other words you fear the probability a future failure may be high.
- **Theoretical Model.** By making assumptions and rules, you can invent a theoretical model to emulate reality. The advantage is that you can compute probabilities of any outcome exactly. For example, we can develop a model of a repeatedly flipped fair coin that lands heads exactly 50% of the time, otherwise tails. Then, we flip the coin 6 times and compute the probability of any outcome $\{TTTTTT, TTTTTH, TTTTHH, \ldots\}$. (You'll do the calculations later.)

Estimating Probabilities from Long-Run Proportions

As an example of method two, we will use `R` to estimate the probability of a fair coin turning up heads. `R` will help us by repeatedly running the coin tossing experiment so we don't have to. In this case, we know that the true probability is 0.5 (it is a fair coin), but let's suppose we don't know that. Instead, we flip a fair coin multiple times and compute the proportion of times the coin turns up 1. That will be our guess of the probability of flipping heads on the next future flip.

```
# create a coin
> coin <- c(0,1) # 0 indicates tails, 1 indicates heads
# flip the coin (run the experiment) 100 times
> sample(x=coin, size=100, replace=T)
  [1] 1 1 0 1 0 0 1 0 1 0 1 0 1 0 0 1 1 1 0 1 1 1 0 1 0 1 0 0 0 0 0 0 1 1 0 1 0
  [38] 0 0 0 1 0 0 0 1 1 0 1 1 1 1 1 0 1 1 0 0 1 1 1 1 0 0 0 0 0 0 1 0 1 1 0 1 1
  [75] 1 1 0 1 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 1
# estimate the proportion of times a head turned up
> sum(sample(x=coin, size=100, replace=T))/100
[1] 0.52
# create an empty vector to hold some results
> p <- NULL
# repeat the estimation 10000 times
> for(i in 1:10000) {
+ p[i] <- sum(sample(x=coin, size=100, replace=T))/100
+ }
# plot a histogram of the results
> hist(p)
```



a **b**

Plot **a** is the $10,000$ estimates of the probability obtained from flippy the coin 100 times. We see that a good fraction of our estimates are below 0.40 and above 0.6. Since the truth is 0.5, our guess at the probability of heads is quite lousy! But, if we run the experiment (coin flipping) $10,000$ times instead of just 100, we get much closer to the truth, as shown by the tightness in histogram **b**.

Thus, we learn that the more information we collect, the more experiments we run, the better we get at predicting a random outcome.

But I bet you already knew that!

R: `sample()`

A note about R's `sample()`. It randomly selects one element in its argument `x`, which is a vector provided by the user. It does this random selection `size` times. If `replace=T`, then it restores the vector before each sampling event. If `replace=F`, then once an element is sampled, it is removed from the vector and will not be sampled again.

For example, I provide vector $(4, 3, 3, 1)$. Suppose `sample()` selects 1 first. Without replacement, the vector then becomes $(4, 3, 3)$. Next, `sample()` selects 3, and the vector becomes $(4, 3)$. Next `sample()`

selects 3. Last, `sample()` selects 4. If `size` is smaller than the length of the vector, `sample()` will stop making random selections after `size` times. When `sample()` is finished, it returns a vector of its selections: $(1, 3, 3, 4)$ in our example.

3 Probability Theory

Definition: *Random Experiment*

A *random experiment* is the chain of circumstances leading up to an *outcome*.

In some cases, the circumstances are tightly controlled by an experimenter and can be repeated many times (e.g. flipping coin, growing corn in a greenhouse, making a compound in chem lab). In other cases, the circumstances are largely out of our control and cannot be repeated (e.g. a plane landing in the Hudson, the election of Obama).

Event

Definition: *Event*

An *event* is a collection of outcomes.

We'll see this more formally later, but for example, the event "Obama wins" contains many election outcomes, including "Obama wins with 270 electoral votes", "Obama wins with 271 electoral votes", "Obama wins with 272 electoral votes", etc. The event a plane crashes contains so many possible outcomes I can't even begin to list them, but one is what happened last Thursday in the Hudson.

3.1 Set Notation

Because *events* are collections of outcomes, we need to review *sets*, which are collections of elements. We will refer to sets as $A, B, C, \ldots$ and elements within sets as $a_1, a_2, \ldots$. To define a particular set, we will write, for example,

$$A = \{a_1, a_2, a_3\}$$

Definition: *universal set or sample space*

The *universal set*, S, is the set of all possible elements.

In the parlance of probability, we name the universal set the *sample space*, S, and define it as "the set of all possible outcomes."

Definition: *subset*

A *subset*, $A \subset B$, means that if $a \in A$, then $a \in B$.

Definition: *null set,* $\varnothing$

The *null set* or *empty set* is the set $\varnothing = \{\}$ containing no outcomes.

Definition: *union*

The union of two sets $A \cup B$ is the set of all elements in A or B, i.e. $a \in A \cup B \Leftrightarrow a \in A$ OR $a \in B$

Definition: *intersection*

The intersection of two sets $A \cap B$ is the set of all elements in A and B, i.e. $a \in A \cap B \Leftrightarrow a \in A$ AND $a \in B$.

Definition: *complement*

The complement of A, written $\bar{A}$ is the set of all elements NOT in A, i.e. $a \in \bar{A} \Leftrightarrow a \notin A$.

Definition: *mutually exclusive*

Two sets are *mutually exclusive* or disjoint if $A \cap B = \varnothing$.

Lemma 1. $A \cup \bar{A} = S$

Proof. To prove this result, you would need to show $A \cup \bar{A} \subset S$ AND $S \subset A \cup \bar{A}$. □

Lemma 2 (distributive laws)**.**

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \end{aligned}$$

Proof. Easiest to prove this result with Venn diagrams. □

Lemma 3 (DeMorgan's Law)**.**

$$\overline{A \cap B} = \bar{A} \cup \bar{B}$$

Proof. First we must show $\overline{A \cap B} \subset \bar{A} \cup \bar{B}$. We will do so by contradiction.
Suppose $\exists a \in \overline{A \cap B}$ for which $a \notin \bar{A} \cup \bar{B}$.
If $a \notin \bar{A} \cup \bar{B}$, then $a \notin \bar{A}$ and $a \notin \bar{B}$ (DeMorgan's logic: $\neg(P \vee Q) = (\neg P) \wedge (\neg Q)$).
If $a \notin \bar{A}$ and $a \notin \bar{B}$, then $a \in A$ and $a \in B$.
If $a \in A$ and $a \in B$, then $a \in A \cap B$, which contradicts our original premise that $a \in \overline{A \cap B}$.
You can prove the other part $\bar{A} \cup \bar{B} \subset \overline{A \cap B}$. □

3.2 The Probability Function

Simple Event

Recall that an *outcome* is the result of a *random experiment*. An *event* is a collection of outcomes. Notice an event $E \subset S$ is a subset of the sample space.

Definition: *simple event*

A *simple event* is an event that cannot be decomposed, i.e. it is one specific outcome of a random experiment.

Lemma 4. *All simple events are mutually exclusive. In other words, if E_i and E_j are different simple events, then $E_i \cap E_j = \varnothing$.*

Proof. Each simple event contains a single outcome, in this case distinct outcomes. We are discussing a single random experiment, so only one outcome can occur. □

Examples: simple events

Example:

- The human genome is made of a long sequence of four different types of nucleotides, called A, C, G, and T. If our random experiment is to select one nucleotide from the human genome, the simple events are that we draw $\{A\}$, $\{C\}$, $\{G\}$, or $\{T\}$.
- If we roll a regular, 6-sided dice, then the simple events are: $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$, and $\{6\}$.
- If you count the number of buses that stop at the bus stop before yours arrives, the simple events are $\{0\}, \{1\}, \{2\}, \ldots$. Though you have faith that your bus will arrive within a finite number of buses, you can't eliminate large counts, so there are infinitely many possible simple events for this case.

Discrete Sample Space

Recall that we defined the *sample space* S for a random experiment as the collection of all possible outcomes of that experiment.

Definition: *discrete sample space*

A discrete sample space is one that contains finite or countably infinite outcomes.

Compound Event

Definition: *compound event*

A *compound event* is a union of 2 or more simple events.

Example:

- Draw a purine from the human genome. The nucleotides are categorized as purines (A and G) and pyrimidines (C and T). So, selecting a purine is the compound event $E = \{A, G\}$.
- Roll an even number is the event $E = \{2, 4, 6\}$.
- Your bus is *not* one of the first two buses, then $E = \{3, 4, 5, \ldots\}$.

Probability

Definition: *probability (THE AXIOMS OF PROBABILITY)*

Probability $P(\cdot)$ is a function that maps events $E \subset S$ to the unit interval $[0, 1]$ such that the following three axioms are satisfed:

1. $P(E) \geq 0$
2. $P(S) = 1$
3. $E_1, E_2, \ldots$ are mutually exclusive, then

$$P(E_1 \cup E_2 \cup \cdots) = \sum_{i=1}^{\infty} P(E_i)$$

Corollary. *Given a probability function $P(\cdot)$ satisfying the probability axioms.*

1. *For a finite set of mutually exclusive events $E_1, E_2, \ldots, E_n$,*
$$P(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_{i=1}^{n} P(E_i)$$
2. *The probability of a compound event is the sum of the probabilities of its simple events. If compound event $E = \{E_1, E_2, \ldots\}$, where all E_i are simple events, then*
$$P(E) = \sum_{i=1}^{\infty} P(E_i)$$
3. $P(\varnothing) = 0$

Examples: probability functions

The axioms of probability are very important for checking whether any function that converts events into numbers can be a probability, but they really don't help you specify specific probability functions for particular situations. We will now consider some reasonable probability functions for some simple random experiments.

- **One-Day Forecast.** Suppose you are a meteorologist and early each morning you need to decide whether it will by sunny, partly sunny, cloudy, or raining today so you know which icon to put on the website. The random experiment is today's weather (lets ignore the possibility of weather that changes during the day). The outcomes are sunny Y, partly cloudy P, cloudy C, and raining R.

  If you study weather in your area over a long period (probability by experience), you observe that 80% of the days are sunny, 10% are partly cloudy, 5% are cloudy and 5% are raining. Today, someone came in to ask you if it will be NOT raining on June 15 (their proposed wedding). You need to compute $P(Y \cup P \cup C)$, but this is just a union of simple events, so $P(Y \cup P \cup C) = P(Y) + P(P) + P(C) = 0.80 + 0.10 + 0.05 = 0.95$. Similarly, you could have noted that the probability they are asking for is $1 - P(R) = 1 - 0.05 = 0.95$.
- **Drawing a diamond.** What is the probability I will draw a diamond from a well-shuffled deck of cards? If you assume all cards are equally likely to be drawn (well-shuffled assumption), then each card (simple event) has probability $\frac{1}{52}$. There are 13 simple events which have the diamond identifier, so the event of drawing a diamond has probability $13 \times \frac{1}{52} = \frac{1}{4}$.

3.3 Constructing Probability Functions

Sample Point Method

Both of the above methods are implementations of the sample point method. The procedure is as follows:

1. First identify all possible outcomes of the random experiment, i.e. identify the sample space S.
2. For each simple event ω, assign a probability $P(\omega)$. Often, such as when throwing a die, drawing a card, flippinng a coin, etc., we will assume that all simple events are equally likely so $P(\omega_i) = P(\omega_j)$ for all simple events $\omega_i \neq \omega_j$. If the sample space is finite of size N, this implies $P(\omega) = \frac{1}{N}$ for all simple events ω.
3. For an arbitrary event E, decompose it into its simple events $E = \{\omega : \omega \text{ is simple event}, \omega \in E\}$, and compute the probability of event E as
$$P(E) = \sum_{\omega \in E} P(\omega)$$

The Importance of Counting

When all simple events have equal probability $P(\omega) = \frac{1}{N}$ for a sample space of finite size N, then we can see that the probability of any event can be written as

$$P(E) = \sum_{\omega \in E} P(\omega) = \sum_{\omega in E} \frac{1}{N} = \frac{1}{N} \sum_{\omega \in E} = \frac{1}{N}|E|$$

where $|E|$ is the size of event E or the number of elements (simple events) in E.

Thus, for computing probabilities in these situations, it becomes important to be able to count the number of outcomes in an event. That is what we tackle next: counting.

3.3.1 mn Rule

mn Rule

Theorem 5. *Given m elements in set $A = \{a_1, \ldots, a_m\}$ and n elements in set $B = \{b_1, \ldots, b_n\}$, then there are mn ways to select one element from set A and one element from set B, e.g. (a_i, b_j).*

Proof. We will use a proof by induction and, though there is no standard notation, we will use $Q_{m,n}$ to represent the number of ways to choose one of m objects and another of n objects.

1. **Show true for $m = 1$.** We want to show that $m = 1$ and arbitrary n, $Q_{1,n} = n$. All pairs are of the form (a_1, b_i) so there exists a one-to-one mapping $f(a_1, b_i) = b_i$ that maps between the pair to an element in B. Since there are n elements in B, we have the result.
2. **Assume true for m.** Assume $Q_{m,n} = mn$.
3. **Show true for $m + 1$.** Consider $A = \{a_1, \ldots, a_m, a_{m+1}\}$. By assumption, there are mn ways to choose pairs where the first element is in the set $A \backslash a_{m+1} = \{a_1, \ldots, a_m\}$. How many pairs involve the element a_{m+1}. Again, there is a mapping $g(a_{m+1}, b_i) = b_i$ between pairs of this type and elements in B. Thus, there are n such pairs and the total is

   $$Q_{m+1,n} = mn + n = (m + 1)n$$

   We can just as easily show the result for fixed m and arbitrary n.

□

Corollary 6. *Consider A_1 with $|A_1| = n_1$ elements, A_2 with $|A_2| = n_2$ elements, …, and A_k with $|A_k| = n_k$ elements. Then the number of k-tuples that can be formed by selecting one element from set A_1, one from A_2, …, and the last from A_k is $n_1 \times n_2 \times \cdots \times n_k$.*

Hint. This rule applies when all elements within the sets (A, B or $A_1, A_2, \ldots$) are distinguishable, but does not distinguish the order of the sets, e.g. it does not distinguish (a_i, b_j) from (b_j, a_i). If you wish to distinguish order, multiply the count by $k!$, the number of ways to order k sets.

Examples

The examples differ from lecture, as several were taken from textbook exercises.

1. **Outfits.** If you have 5 trousers and 4 shirts, how many unique outfits can you put together?

   $$5 \times 4 = 20$$

2. **Telephone numbers.** How many 7-digit telephone numbers are there if numbers starting with 555 are eliminated?

   The number of 3-digit numbers, excluding 555, is

   $$10 \times 10 \times 10 - 1 = 999$$

   The number of 4-digit numbers is

   $$10 \times 10 \times 10 \times 10 = 10^4 = 10,000$$

   The number of ways to put them together into 7-digit numbers $XYZ - ABCD$, where XYZ is from the 3-digit set and $ABCD$ is from the 4-digit set is

   $$999 \times 10,000 = 9,990,000$$

3. **The random brain. Can human brains generate truly random numbers?** Suppose a research subject is asked to generate a sequence of random numbers, where each number is supposed to be chosen completely at random from the set $N_9 = \{0, 1, \ldots, 9\}$. In the sequence of numbers the subject generates, you count the number of times (1) sequential numbers appear side-by-side (e.g. $(2, 3)$) and (2) the number of times two identical numbers appear side-by-side (e.g. $(4, 4)$). You repeat the experiment for many different subjects and compile the following data (from Bains, W. (2008) Random number generation and creativity. *Medical Hypotheses*. **70**: 186-190.):

	Probability	
	Observed	Expected (computed in what follows)
$N_\pm = \{(n, n \pm 1) : n, n \pm 1 \in N_9\}$	0.35	$P(N_\pm) = 0.18$
$N_n = \{(n, n) : n \in N_9\}$	0.04	$P(N_n) = 0.10$

   The random experiment in this question is the generation of a pair of integers from N_9 by a human brain. The sample space $S = \{(i, j) : i, j \in \{0, 1, \ldots, 9\}\}$, and all outcomes should be equally likely if the hypothesis of random generation is true. To answer the question, we need to compute the probability of each of the events defined as $N_\pm$ and N_n, but because all simple events are equally likely, we have

   $$P(N_\pm) = \frac{|N_\pm|}{|S|} \qquad\qquad P(N_n) = \frac{|N_n|}{|S|}$$

   The number of possible pairs in the sample space S is $|S| = 10 \times 10 = 100$.

   The number of pairs in set $N_\pm$ is $|N_\pm| = 2 \times 1 + 8 \times 2$, where we recognize that the second set (B in the parlance of Theorem 5) depends on the choice of the first element. So, we have $A = \{0\}$ and $B = \{1\}$ or $A = \{1\}$ and $B = \{0, 2\}$, etc. Sum them all up to get the total count.

   The number of pairs in set N_n is $|N_n| = 10$.

   We conclude $P(N_\pm) = \frac{18}{100} = 0.18$ and $P(N_n) = \frac{10}{100} = 0.1$. The observed probabilities are quite different. We would of course need to account for sampling error in the calculation of observed probabilities when making actual inference.

4. **Birthdays.** What is the probability that no two people in a group of 24 (the size of your birthday dataset) share the same birthday?

   Suppose that all days (excluding leap day) are equally likely to be a person's birthday. Let X_i be the birthday (in days out of 365) of the ith individual. The number of datasets $(X_1, \ldots, X_{24})$ possible is

   $$365^{24} \approx 3.126286 \times 10^{61}$$

   The event that no two people share the same birthday is the event that all birthdays are unique. The number of datasets with this property is

   $$365 \times 364 \times \cdots \times (365 - 24 + 1) \approx 1.443268 \times 10^{61}$$

There are 365 choices for X_1, but only 364 choices for X_2 since $X_2 \neq X_1$ is required, and so on for the result. The probability of the event is therefore

$$\frac{1.443268 \times 10^{61}}{3.126286 \times 10^{61}} \approx 0.4616557.$$

3.3.2 Permutations

Permutation

Definition: *permutation*

A permutation is an ordered arrangement of r objects. When we select those r objects from n possible objects, then the number of permutations is P_r^n.

Theorem 7.

$$P_r^n = \frac{n!}{(n-r)!}$$

Proof. Use the mn rule, as we demonstrated for the birthday example, where we had to find the number of ways to choose 24 birthdays from the list $\{1, 2, \ldots, 365\}$. □

R Hint: To compute P_r^n in `R`, try `choose(n,r)*factorial(r)`. The function `choose(n,r)` is C_r^n (defined later) and the function `factorial(r)` returns $r!$.

Examples

1. **Candy.** If you have a bag of candies with 6 different colors, how many ways can you select 3 distinct colors, where order matters, so red, white, blue is different from red, blue, white?

$$P_3^6 = 6 \times 5 \times 4 = 120$$

2. **Campaigning.** You were hired to arrange McCain's campaign schedule. He needed to visit 6 distinct cities. How many ways could you have arranged the trips for him?

$$P_6^6 = 6! = 720$$

3. **Cards.** How many hands of size 5 can you deal from a deck of 52 cards?

$$P_5^{52} = 311875200$$

However, note these are ordered hands, and you probably don't care whether your were dealt the King first or last.

3.3.3 Combination

Combination

When you want to know the number of ways to arrange unordered objects, you use combination instead.

Theorem 8. *The number of ways to partition n distinct objects into k distinct groups such that the groups contain $n_1, n_2, \ldots, n_k$ objects such that each object appears in only one group and $\sum_{i=1}^{k} n_i = n$ is*

$$C_{n_1,\ldots,n_k}^n = \binom{n}{n_1 \quad n_2 \quad \cdots \quad n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$$

When $k = 2$, let $n_1 = r$ and $n_2 = n - r$. Then, the notation becomes simpler, and we have

$$C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{P_r^n}{r!}$$

Notes:

1. $C^n_{n_1,\ldots,n_k}$ is called the *multinomial coefficient* because it is the constant coefficient of $x_1^{n_1}x_2^{n_2}\cdots x_k^{n_k}$ in the expansion of $(x_1 + x_2 + \cdots + x_k)^n$. Similarly, C^n_r is called the *binomial coefficient* because it is the coefficient of $x^r y^{n-r}$ in the expansion of $(x + y)^n$.

2. A contribution from the class: $C^n_{n_1,\ldots,n_k} = C^n_{n_1} \times C^{n-n_1}_{n_2} \times \cdots \times C^{n-n_1-\cdots-n_{k+1}}_{n_k}$, so you can use the following `R` code to compute multinomial coefficients:

   ```
   choose(n,n1)*choose(n-n1,n2)*choose(n-n1-n2,n3)*···*choose(nk,nk)
   ```

Examples

1. **Taxis.** You need to dispatch 9 taxis to three airports, and you would like 3 to go to airport A, 5 to airport B, and 1 to airport C. If one taxi is in need of repair, what is the probability that it goes to airport C?

   There are $C^9_{3,5,1} = \frac{9!}{3!5!1!} = 504$ ways to dispatch the taxis as requested and without regard to the broken taxi.

   The number of ways to dispatch the taxis such that the desired event occurs is $C^8_{3,5} = 56$ because the broken taxi must go to airport C (no randomness there) and the remaining 8 taxis can be sent in any combination that yields 3 to A and 5 to B. Overall, the probability of dispatching the broken taxi to C is $56/504 = 1/9$.

   Another contribution from the class: you can think about it another way. How many ways can airport C select it's taxi? Without regard to broken taxis, airport C has 9 choices. Under the desired event, airport C can only choose the broken taxi (1 choice), so the probability of the target event is $1/9$.

   As this problem demonstrates, there is more than one way to think about a problem and derive an answer. Here, we changed perspective from distributing taxis to selecting taxis, so the random experiment changed and the sample space did as well (it shrunk).

2. **Cards Revisited.** How many hands of size 5 can you distribute from a deck of 52 cards?

   Recall, we though previously that order doesn't matter, and now we have the tools to make the calculations:

$$\binom{52}{5} = \frac{52!}{5!47!} = 2598960$$

There is more than one way to solve a combinatorics problem, which makes it easy to fool yourself into thinking your logic is right when it isn't. The only way to learn the techniques and avoid the mistakes is to practice, practice, *practice*.

3.3.4 Examples Galore

Example: throwing a die

Suppose you throw a fair die 6 times. What is the probability that you see every face exactly once?

Denominator. Count the number of outcomes resulting from throwing a die 6 times. 6 choices for each throw results in 6^6.

Numerator. We will see every face only if we observe each one exactly once. They can be arranged in any order, and there are $P_6^6 = 6!$ orders.

Therefore, the probability of observing every face on the first 6 throws is

$$\frac{6!}{6^6} \approx 0.0154321$$

Example: throwing a die II

Suppose you throw a fair die 4 times. What is the probability that you see 4 distinct faces?

Numerator. There are 6 faces $\{1, 2, 3, 4, 5, 6\}$. You can choose any 4 of them with no repeats. The number of ordered ways to do that is $P_4^6 = \frac{6!}{2!}$.

The probability is

$$\frac{6!}{2!6^4} \approx 0.2777778$$

Example: throwing a die III

Suppose your die has two faces with 5, and is otherwise normal. What is the probability that in the first 5 throws you see all numbers in the set $\{1, 2, 3, 4, 5\}$.

Numerator. The 5 is the trouble, so let's count the ways we can get a 5 first. The 5 is distinguished by the throw it shows up and which of the two faces with 5 is actually showing. The mn rule applies with 5 throws and 2 faces, so there are 10 ways to place a 5 in the sequence. After the 5 is placed, there are 4 remaining slots into which we have to place the remaining 4 faces. That event can happen in $4!$ ways, as we calculated in the first example.

The probability is

$$\frac{5 \times 2 \times 4!}{6^5} \approx 0.0308642$$

Example: distributing jobs

Suppose I have n compute jobs to distribute to N computers. I decide to distribute each job randomly to any one of the N computers. What is the probability that k or more jobs are distributed to computer A.

Let P_i be the probability that exactly i jobs are distributed to the computer in question. Then, the probability we seek is

$$\sum_{i=k}^{n} P_i$$

Now, for computing P_i, we need to count.

Denominator. The number of ways to distribute n jobs to N computers in the manner described is N^n, N choices for job 1, N choices for job 2, etc.

Numerator. The number of ways to distribute i jobs to computer A and the remaining $n - i$ jobs to any of the other computers is $\binom{n}{i}(N-1)^{n-i}$. There are $\binom{n}{i}$ ways to assign i of the jobs to computer A. The other $n - i$ jobs can be assigned to any of the $N - 1$ remaining computers, much like the calculation used for the denominator.

The probability, therefore is

$$P_i = \frac{n!(N-1)^{n-i}}{i!(n-i)!N^n}$$

3.4 Conditional Probability

Example: Testing

Suppose you observe the following results for a sample of 1000 people, 500 with a disease (event D) and 500 without the disease (event H). Each person is given a test for the disease. Out of 1000 tests, 492 test positive (event P) and 508 test negative (event N). Each of the 1000 people tested falls into one of four categories, and the exact counts are shown below.

	Test Positive P	Test Negative N	Row Totals
Person has disease D	487	13	500
Person is healthy H	5	495	500
Column Totals	492	508	1000

Consider the random experiment of drawing one of the 1000 people in the sample at random and then noting their disease and test status. There are four possible outcomes and the probability that the random individual falls into any one is given by the count of that category over 1000, so

$$\begin{aligned} P(D \cap P) &= \frac{487}{1000} = 0.487 \\ P(D \cap N) &= \frac{13}{1000} = 0.013 \\ P(H \cap P) &= \frac{5}{1000} = 0.005 \\ P(H \cap N) &= \frac{495}{1000} = 0.495 \end{aligned}$$

In addition, the probability of any of the four events can be computed using the row and column totals, so

$$\begin{aligned} P(D) = P(H) &= \frac{500}{1000} = 0.5 \\ P(P) &= \frac{492}{1000} = 0.492 \\ P(N) &= \frac{508}{1000} = 0.508 \end{aligned}$$

There are many interesing questions we can ask with regard to this data.

- What is the probability of testing positive *if the patient has the disease*?
- What is the probability of having the disease *given you test positive*?
- What is the probability of having the disease *if the patient tests negative*?

In each case above, the phrase in italics identifies a random event, and the questions are stated in such a way that these events are assumed to be true.

If a particular event, say A, is imposed, we are redefining the random experiment. Instead of considering the random experiment with outcomes on the entire sample space S, now only events falling in $A \subset S$ are considered. Experiments ending with outcomes in $\bar{A}$ are ignored. Within the subset of experiments with outcomes in A, we then ask about the probability of an additional event B, such as the event of testing positive P.

Conditional Probability

Definition: *conditional probability*

> For events A and B, the probability of event A conditional on event B is the *conditional probability*
>
> $$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$
>
> when $P(B) > 0$.

Example: labrador retriever

Labrador retrievers have a gene that controls their coat color. There are two flavors of this gene B is called the black gene, and b is called the chocolate gene. Like humans, dogs have two copies of every gene (the combination is called their genotype), but all they need is one copy of the black gene to have the black coat color, as summarized in this table.

	Genotype		
	BB	Bb	bb
Coat color:	black	black	chocolate

When dogs breed, each gene is equally likely to be passed to the offspring. The dad passed on one of his genes, the mom passes on one of her genes, and the puppy ends up with two genes. If two dogs with genotypes Bb are crossed four types of puppies can result.

		Gene Passed by Father	
		B	b
Gene Passed by Mother	B	BB	Bb
	b	bB	bb

For now, I'll tell you all 4 possibilities are equally likely.

What is the conditional probability that the puppy is black given that the father passed his b gene to the puppy? If the puppy is black, then mom must have given him a B, so the probability we seek is:

$$P(\text{puppy black} \mid \text{dad passed } b) = \frac{P(\text{puppy black} \cap \text{dad passed} b)}{P(\text{dad passed } b)} = \frac{P(Bb)}{P(\text{father passed } b)} = \frac{1/4}{1/2} = \frac{1}{2}$$

3.5 Independence

Independence

Definition: *independence*

> Two events A and B are said to be *independent* if one of the following holds
>
> 1. $P(A \mid B) = P(A)P(B)$
> 2. $P(B \mid A) = P(B)$
> 3. $P(A \mid B) = P(A)$

Corollary 9. *The definition implies the three properties are equivalent definitions of independence.*

Proof. To show (3) $\Rightarrow$ (1), notice $P(A \cap B) = P(A \mid B)P(B)$ by definition of conditional probability.

But (3) implies $P(A \mid B)P(B) = P(A)P(B)$, which shows (1).

Other are shown in a similar way. □

Examples: independence

1. **Labradors I.** It turns out that moms and dads pass their genes independently. In other words, if the mom is passing a B, it has no impact on whether the dad will also pass a B. Let B_m be the event that mom passes B. Let b_m be the event that mom passes b. Similarly, we define B_d and b_d. The independence assumption implies

   $$P(B_m B_d) = P(B_m)P(B_d) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

   Repeat for the other outcomes, and we have now justified the assumption that all puppy genotypes are equally likely.

2. **Labradors II.** Some labs can be yellow. There is another gene that we must consider to understand how that happens. This gene has two varieties also: E and e, the latter is called the yellow gene because if a lab has two copies, it will be yellow. Suppose our parents both have genotypes $BbEe$, then what is the probability of a $bbee$ puppy if both genes are transmitted to offspring randomly?

   We know $P(bb) = P(ee) = \frac{1}{4}$. By independence, then $P(bbee) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$.

3.6 Trump Example

Trump Example

A card game is played with 52 cards divided equally between four players, North, South, East and West, all arrangements being equally likely. Thirteen of the cards are referred to as trumps. If you know that North and South have 10 trumps between them, what is the probability that all three remaining trumps are in the same hand?

$$\frac{\binom{2}{1}\binom{23}{10}}{\binom{26}{13}} = 0.22$$

The first thing to notice is that the problem is asking for a conditional probability, where the condition is that two teams, North and South, are known to have received 10 trump cards. There are many ways to distribute 10 of 13 trump cards, and we have tools for counting the ways, but because these 10 were distributed already as the condition, we don't need to count them. We take this fact as given. Also, 13 of the 49 non-trump cards have been distributed to North and South. There are many ways to do that, but again it is part of the condition and we'll ignore it.

What we are left with is 26 cards, 3 trumps and 23 non-trumps, that need to be distributed, 13 to East and 13 to West. There are $\binom{26}{13}$ ways to make this distribution with no restrictions. That's the denominator.

For the numerator, we have to choose one of East or West to receive the three trumps and we need to choose one way of distributing 10 more non-trumps to the chosen one. The mn rule applies, where $m = \binom{2}{1}$, the number of ways to choose one from $\{\text{East}, \text{West}\}$, and $n = \binom{23}{10}$, the number of ways to select 10 of 23 cards to give to the chosen one.

Suppose I asked instead what is the probability that 2 trumps ended up in one hand and 1 ended up in the other. The answer is

$$\frac{\binom{2}{1}\binom{3}{2}\binom{23}{11}}{\binom{26}{13}} = 0.78$$

Here, we have an applicaton of the mn-rule for three groups. We have to select East or West, with $\binom{2}{1}$ ways to do that. We have to select one way to distribute 2 trumps to the chosen hand and 1 to the other, and there are $\binom{3}{2}$ ways to do that. Finally, we have to select one way to distribute 11 of the 23 non-trumps to the chosen hand, leaving 12 for the other hand.

Using `R` to Check Yourself

As you saw, the above question is not easy to answer. When you are confronted with such problems, you can check your calculations using `R`. While `R` can never prove your answer is correct, it can certainly indicate when you are wrong, and at least lend support when you are correct.

The following `R` code defines a function `prob()` that can be used to run the experiment an arbitrary `n` times. It then reports the proportion of times where all three trumps ended in one hand or two in one hand, one in the other. (Actually, I've spruced up the function since class to handle the second problem also. See if you can follow.)

```
# create the 26 cards that need to be distributed to East and West
cards <- c(rep(1,3), rep(0,23))
# define the function
prob <- function(n=10000, cards, match=c(0,3)) {
   # will hold count of matching outcomes
   m <- 0
   # iterate the requested number of times
   for(i in 1:n) {
      # count the number of trump cards dealt to the player
      s <- sum(sample(x=cards, size=13))
      event.matched <- 0
      # check whether the number of cards matches any of the values in match argument
      for(j in 1:length(match)) {
         if(match[j] == s) {
            event.matched <- 1
            break;
         }
      }
      # increment count of matching events
      if(event.matched) {
         m <- m + 1
      }
   }
   m/n # the results of this calculation are returned from the function
}
prob(n=100000, cards=cards)
prob(n=100000, cards=cards, match=c(1,2))
```

The output is

```
[1] 0.22007
[1] 0.7795
```

which is certainly suggestive, and definitely not contradictory of our analytic calculations.

3.7 Multiplication Law

Multiplication Law

Theorem 10. *For any two events A and B*

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

If A and B are independent, then

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

Proof. Trivial application of conditional probability definition or independence definition. □

Corollary 11.

$$\begin{aligned} P(A \cap B \cap C) &= P(A)P(B \mid A)P(C \mid A \cap B) \\ &= P(A)P(C \mid A)P(B \mid A \cap C) \\ &= P(B)P(A \mid B)P(C \mid A \cap B) \\ &= P(B)P(C \mid B)P(A \mid C \cap B) \\ &= P(C)P(A \mid C)P(B \mid A \cap C) \\ &= P(C)P(B \mid C)P(A \mid B \cap C) \end{aligned}$$

Proof. All are proved in the same way, so we'll do the first. Suppose $D = A \cap B$. Then,

$$\begin{aligned} P(A \cap B \cap C) &= P(D \cap C) \\ &= P(D)P(C \mid D) \\ &= P(A \cap B)P(C \mid A \cap B) \\ &= P(A)P(B \mid A)P(C \mid A \cap B) \end{aligned}$$

□

Advice. Let's rearrange the last expression to yield

$$\begin{aligned} P(C \mid A \cap B) &= \frac{P(C \cap B \cap A)}{P(B \mid A)P(A)} \\ &= \frac{P(C \cap B \mid A)P(A)}{P(B \mid A)P(A)} \\ P(C \mid A \cap B) &= \frac{P(C \cap B \mid A)}{P(B \mid A)} \end{aligned}$$

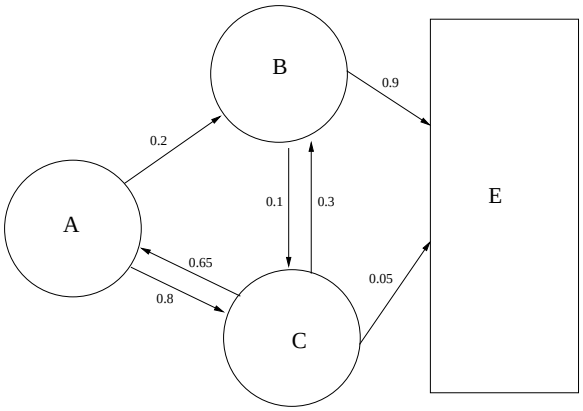
This last relationship demonstrates application of the conditional probability rule when the condition A is universal, i.e. assumed on both the left and the right. I recommend you get *very* familiar with how to move events across that condition bar | using these kinds of relationships. Practice, practice until it becomes second nature, just as second nature as FOIL: $(a + b) \times (c + d) = ac + ad + bc + bd$.

Corollary 12. *For any events* $A_1, A_2, \ldots, A_k$,

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = P(A_1)P(A_2 \mid A_1) \cdots P(A_{k-1} \mid A_1, A_2, \ldots, A_{k-2})P(A_k \mid A_1, A_2, \ldots, A_{k-1})$$

Proof. Repeated application or induction of rule for $P(A \cap B)$. □

Example: Multiplication Rule



Suppose you are confronted with the above cave system, and you start in room A. Each arrow indicates where you might walk next and the probability with which you would make that move given you start in the room from which the arrow emanates. For example, from room A, you have a 20% chance of walking to room B and an 80% chance of walking to room C. Let X_i be the event that you walk into room X on your ith move. If E is the exit, what is the probability of event $B_1 \cap E_2$, i.e. you move from room A to B and then from B directly to the exit E?

$$P(B_1 \cap E_2) = P(E_2 \mid B_1)P(B_1) = 0.9 \times 0.2 = 0.18$$

What about the event $B_1 \cap C_2 \cap E_3$?

$$\begin{aligned} P(B_1 \cap C_2 \cap E_3) &= P(B_1)P(C_2 \mid B_1)P(E_3 \mid B_1 \cap C_2) \\ &= P(B_1)P(C_2 \mid B_1)P(E_3 \mid C_2) \\ &= 0.2 \times 0.1 \times 0.05 = 0.001 \end{aligned}$$

Notice the simplification in step 2 above. We see that once you are in room C_2, the fact that you were just previously in B_1 has no impact on where you will move next. In other words, where you move next depends only on your current location, not any previous location. *Next room is **independent** of all previous rooms but the current room.*

3.8 Addition Rule

Addition Rule for 2 Events

Theorem 13. *For any events A and B,*

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

Proof. In the following Venn diagram, event A is shown in blue, event B in pink, and the sample space in white. The intersection $A \cap B$ appears as purple. The union $A \cup B$ is the colored region.

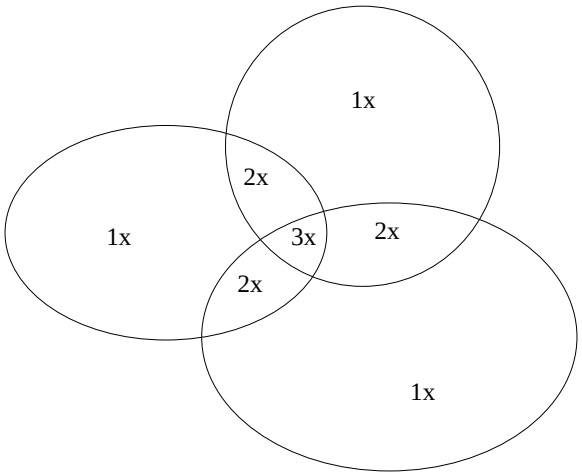
One need only check the Venn diagram to see that if we use $P(A \cup B) = P(A) + P(B)$, then we have counted the purple region $(A \cap B)$ twice. Thus, we need to subtract $P(A \cap B)$ once to get the probability of the colored region. □

Addition Rule for 3 Events

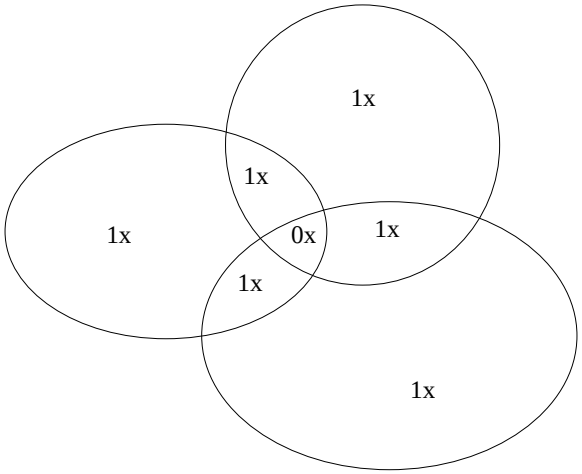
Theorem 14. *For any events A, B, and C,*

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

Proof. After adding the first three terms, the Venn diagram has counted each of the 7 regions in the figure as follows:



After subtracting the next three terms, the Venn diagram becomes:



Since the center region is not counted, we need to add it in with the last term $P(A \cap B \cap C)$. □

General Addition Rule

Theorem 15. *For any events* $A_1, \ldots, A_n$

$$P\left(\cup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n)$$

Proof. We shall not go through the torture, but I think you can see the pattern from 2 and 3 events. □

Example: addition rule

1. Draw one random card from a well-shuffled deck. Let A be the event that the card drawn is an ace. Let K be the event that the card is a King. We apply the addition rule for two events and notice that $A \cap K = \varnothing$.

$$\begin{aligned} P(A \cup K) &= P(A) + P(K) - P(A \cap K) \\ &= \frac{4}{52} + \frac{4}{52} - 0 \\ &= \frac{2}{13} \end{aligned}$$

2. Let T be the event that the card is a 2. Let R be the even that the card is red.

$$\begin{aligned} P(T \cup R) &= P(T) + P(R) - P(T \cap R) \\ &= \frac{4}{52} + \frac{26}{52} - \frac{2}{52} \\ &= \frac{28}{52} \end{aligned}$$

3. Throw two dice and sum the numbers that appear. Let E be the event that the sum is even. Let O be the event that the sum is odd. E and O are mutually exclusive and clearly since any number must be even or odd $E \cup O = S$. Thus, by the probability axioms

$$P(E \cup O) = P(S) = 1$$

4. Let M_2 be the event that the sum of two dice is a multiple of 2 and M_3 be the event that the sum is a multiple of 3.

$$\begin{aligned} M_2 &= \{(1,1),(1,3),(3,1),(2,2),(1,5),(5,1),(2,4),(4,2),(3,3),(2,6),(6,2),(3,5), \\ &\quad (5,3),(4,4),(4,6),(6,4),(5,5),(6,6)\} \\ M_3 &= \{(1,2),(2,1),(1,5),(5,1),(2,4),(4,2),(3,3),(3,6),(6,3),(4,5),(5,4),(6,6)\} \\ P(M_2 \cup M_3) &= P(M_2) + P(M_3) - P(M_2 \cap M_3) \\ &= \frac{18}{36} + \frac{12}{36} - \frac{6}{36} \\ &= \frac{24}{36} = \frac{2}{3} \end{aligned}$$

Complements

Corollary 16. *For any event A,*

$$P(\bar{A}) = 1 - P(A)$$

Proof. Notice that $A \cup \bar{A} = S$. Further, $A \cap \bar{A} = \varnothing$, so any event A and its complement are mutually exclusive. Therefore, the traditional probability rule and probability axioms apply to give us

$$1 = P(S) = P(A \cup \bar{A}) = P(A) + P(\bar{A})$$

and rearrangement gives us the result. □

Example:

What is the probability that an odd number shows up when throwing two dice (event O)?

We have already computed $P(E) = \frac{18}{36}$, so $P(O) = 1 - P(E) = \frac{18}{36}$.

More Complements

Corollary 17. *For any events A and B,*

$$P(A \cap \bar{B}) = P(A) - P(A \cap B)$$

Proof. Real easy by Venn diagram, but also

$$\begin{aligned} P(A \cap \bar{B}) &= P(A)P(\bar{B} \mid A) \\ &= P(A)\left[1 - P(B \mid A)\right] \\ &= P(A) - P(A)\frac{P(B \cap A)}{P(A)} \\ &= P(A) - P(B \cap A) \end{aligned}$$

□

Notice, $P(\bar{B} \mid A) = 1 - P(B \mid A)$, as per corrollary 16 applied to the conditional probability function $P(\cdot \mid A)$.

Example:

Suppose I is the event that a person contracts disease I sometime in their life. Suppose II is the event that someone contracts disease II in their life. You are given

$$\begin{aligned} P(I) &= 0.10 \\ P(II) &= 0.15 \\ P(I \cap II) &= 0.03 \end{aligned}$$

Compute the following,

- What is $P(I \cup II)$? There's a union. Think addition rule.

$$\begin{aligned} P(I \cup II) &= P(I) + P(II) - P(I \cap II) \\ &= 0.10 + 0.15 - 0.03 = 0.22 \end{aligned}$$

- What is $P(I \cap \bar{II})$? There's an intersection. Think multiplication rule. Except, wait. There's a complement, sothere is a shortcut to the result.

$$\begin{aligned} P(I \cap \bar{I}I) &= P(I) - P(I \cap II) \\ &= 0.10 - 0.03 = 0.07 \end{aligned}$$

- What is $P(I \cap II \mid I \cup II)$? There's a condition. Think definition of conditional probability.

$$\begin{aligned} P(I \cap II \mid I \cup II) &= \frac{P[(I \cap II) \cap (I \cup II)]}{P(I \cup II)} \\ &= \frac{P(I \cap II)}{P(I \cup II)} \\ &= \frac{0.03}{0.22} \approx 0.136 \end{aligned}$$

  Notice, the distribution laws were used to find the event in the numerator, so $(I \cap II) \cap (I \cup II) = ((I \cap II) \cap I) \cup ((I \cap II) \cap II) = (I \cap II) \cup (I \cap II) = I \cap II$.

3.9 A Procedure for Calculating Probabilities

Everyone finds it challenging to solve word problems that ask for probabilities of events. If only there were a regular procedure that could be followed to reliably come up with an answer. Unfortunately, no one specific procedure works for all real-life cases. Here is an outline of a procedure that you may find useful to provide a scaffold to follow when you are stuck. You need to fill in the many missing details.

1. **What is the experiment?**
2. **What is the sample space?** Are the outcomes equally likely?
3. **Define events.** Identify and name events relevant to the question. In particular, identify the target event, that is the event for which you are asked to compute a probability. Using set theory, write the target event as a composition of other events. The goal is to get the target event on the left, and an expression of other events whose probabilities are known on the right. Sounds simple. Can be hard.
4. **Apply laws.** Apply the probability function to both sides and use probability laws to compute the probabilities. Sounds simple. Can be tricky.

The process of applying the rules can be *iterative*, so that you may start wrong, but correct yourself by returning to step 1 after figuring out some details in later steps. Also, there is more than one way to apply the steps.

Let's apply it to an example.

Two Boys

A family has two children. What is the probability that both are boys given at least one is a boy?

1. Pick a random family with two children and record the sexes of the children.
2. $S = \{GG, GB, BG, BB\}$ and we notice all outcomes are equally likely.
3. Two events are discussed in the problem statement.
   - Let E be the event that at least one child is a boy. Clearly, $E = \{GB, BG, BB\}$.
   - Let B be the event that both are boths. $B = \{BB\}$.

   Furthermore, it is $B \mid E$ (an abuse of notation, but hopefully communicates what we want) that is the target event.
4. Apply probability function, notice condition, and apply definition of conditional probability.

$$P(B \mid E) = \frac{P(B \cap E)}{P(E)}$$

   But $B \cap E = \{BB\}$ and because of equally likely outcomes, $P(B \mid E) = \frac{1}{3}$.

A sidetrack. We might have been tempted to define the experiment as "Pick a random family and record the number of boys", so the sample space is $S = \{0, 1, 2\}$, but you should notice that outcome 1 is not a simple event. There are two outcomes GB and BG that result in 1. In fact, the number of boys is a *random variable*, something we will define in the near future. You should distinguish: (1) processed outcomes, such as $\{0, 1, 2\}$, and (2) direct outcomes of random experiments. You want unprocessed outcomes for this procedure.

Two Boys, Take 2

What is the probability that both are boys given the first is a boy?

1. Same.
2. Same.
3. Again, there are two events.

- Let F be the event that the first child is a boy. $F = \{BG, BB\}$.
- B is unchanged.

And again, we seek $B \mid F$.

4. $P(B \mid F) = \frac{1}{2}$, again by law of conditional probability.

3.10 Law of Total Probability

Partition

Definition: *partition*

The events $B_1, \ldots, B_n$ form a partition of S if

1. they are mutually exclusive $B_i \cap B_j = \varnothing$ for all $i \neq j$, and
2. and exhaustive $\cup_{i=1}^{n} B_i = B_1 \cup \cdots \cup B_n = S$

Properties:

1. Each simple event belongs to exactly one set in the partition. Suppose it belongs to two, then they would not be mutually exclusive.
2. Simple events (outcomes) form (largest n) partition of the sample space.
3. Any set A can form a partition when combined with its complement $\bar{A}$.

Law of Total Probability (LTP)

Lemma 18 (Law of Total Probability (LTP))**.** *Given events A and B,*

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \bar{B}) \\ P(A) &= P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B}) \end{aligned}$$

More generally, for a partition $B_1, \ldots, B_n$ of S,

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A \cap B_i) \\ P(A) &= \sum_{i=1}^{n} P(A \mid B_i)P(B_i) \end{aligned}$$

Proof. You were asked to generate this proof in HW2 for $n = 3$. You must prove two things

1. $A = (A \cap B_1) \cup \cdots \cup (A \cap B_n)$
2. $(A \cap B_i) \cap (A \cap B_j) = \varnothing$ for all $i \neq j$ (mutually exclusive)

Then, the first version of LTP follows as an application of the axiom of probability #3, and the second follows after applying multiplication law.

To show 1:

- First show $\cup_{i=1}^{n} A \cap B_i \subset A$. Suppose $\omega \in \cup_{i=1}^{n} A \cap B_i$, then $\omega \in A \cap B_i$ for some i, which implies $\omega \in A$.

- Then show $A \subset \cup_{i=1}^{n} A \cap B_i$. Suppose $\omega \in A$ but $\omega \notin \cup_{i=1}^{n} A \cap B_i$. The latter implies $\omega \notin A \cap B_i$ for all i. But $\omega \in A$, implies $A \notin B_i$ for all i, which contradicts that B_i is a partition.

To show 2: $(A \cap B_i) \cap (A \cap B_j) = A \cap B_i \cap B_j = \varnothing$ because $B_i \cap B_j = \varnothing$. (uses intersection $\cap$ is commutative, i.e. $A \cap B = B \cap A$)

□

Examples: LTP

Example:

Suppose two factories, I and II, produce widgets, with I producing twice as many as II. $1/5$ of widgets produced by factory I are defective, and $1/20$ of widgets produced by factory II are defective. What is the probability that a random widget is defective?

1. Select a random widget.
2. Each widget is produced by one factory and is either defective or not.

   $$S = \{(0, I), (1, I), (0, II), (1, II)\}$$

   where each pair indicates defective status (0 or 1) and factory of origin. There are many more facts about each widget, for example perhaps each is a particular weight or color, but none of those details are mentioned or matter to the problem. All the details that are mentioned are included in my sample space.
3. There are several events associated with this problem.
   - D: widget is defective, also target event
   - $\bar{D}$: widget is not defective
   - I: widget made in factory I
   - II: widget made in factory II

   My target is D. I could write some relations between D and the other events, but let's skip this and see if something jumps out at us in the next step.
4. I seek $P(D)$. Let's see what I know.
   - $P(I) = 2/3$, $P(II) = 1/3$.
   - $P(D \mid I) = 1/5$, $P(D \mid II) = 1/20$.

   I know a lot and in fact, *if only I knew which factory the widget came from*, I could answer the question. This is a key phrase. If you find yourself saying "if only I knew...", then think LTP. Clearly $\{I, II\}$ form a partition of the sample space, then applying LTP, we have

   $$P(D) = P(D \mid I)P(I) + P(D \mid II)P(II) = \frac{1}{3} \times \frac{2}{3} + \frac{1}{20} \times \frac{1}{3} = \frac{43}{180}$$

3.11 Bayes' Rule

Bayes' Rule

The Presbyterian minister, Thomas Bayes, lived three centuries ago. He was concerned with computing "inverse probabilities", in a sense probabilities on events taken out of order. At the time, probabilists were really good at computing forward probabilities:

- In the cave example, what is the probability that I make it out in as few steps as possible?
- If there are 5 red candies and 2 blue candies in a bag, what is the probability the first 3 candies I draw are red?

- If the stock market goes up today, what is the probability it will go up tomorrow?

Inverse probabilities were another story. Examples include

- What is the probability that I was in room C *before* I walked outside?
- If I have drawn three red candies out of a bag, can I say anything about the color of candies in the bag?
- If the stock market went up today, what is the probability it was up yesterday?

His simple (and you will see how simple in a second) result has had an amazing impact on the field of statistics. You can now take undergraduate courses in "Bayesian Statistics".

Lemma 19. *Given a partition $B_1, B_2, \ldots, B_n$ of S such that $P(B_i) > 0$ for all i, then for any event A*

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}$$

Proof. Trivial application of definition of conditional probability, multiplication law, and LTP.

$$\begin{array}{rcll} P(B_j \mid A) & = & \frac{P(B_j \cap A)}{P(A)} & \text{definition of conditional probability} \\ & = & \frac{P(A|B_j)P(B_j)}{P(A)} & \text{multiplication law} \\ & = & \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} & \text{LTP} \end{array}$$

□

So simple, yet so powerful.

Example:

> You are trying to fill a job. Suppose the eastern branch of your company produces a name of its best candidates, including 5 women and 2 men. Meanwhile, the western branch, which you know produces far better candidates, produces a list with 2 women and 6 men.
>
> You need a list with 9 total candidates, but you trust the western branch to produce better candidates. You decide to randomly choose one candidate from the eastern list and add it to the western list. Then, you put the 9 names in a jar and randomly draw one name to fill the job.
>
> What is the probability that you drew one of the 5 women from the eastern list when augmenting the western list *given* that you ultimately gave the job to a male?
>
> 1. The experiment is two-fold. Draw one from the eastern list and draw one from the augmented west list. In both cases, note the sex.
> 2. $S = \{FF, FM, MF, MM\}$
> 3. Let F_1 be the event that the first draw is female, M_1 that it is male. Similarly, let F_2 be the event that the final draw is female, else M_2. $F_1 \mid M_2$ is our target.
> 4. We seek $P(F_1 \mid M_2)$, but this is an inverse probability because it asks us to compute what happened in an earlier step given the final outcome. Think Bayes' Rule!
>
> $$P(F_1 \mid M_2) = \frac{P(M_2 \mid F_1)P(F_1)}{P(M_2 \mid F_1)P(F_1) + P(M_2 \mid M_1)P(M_1)}$$
>
> Let's now work out the various probabilities. $P(F_1) = 2/7$ and $P(M_1) = 1 - 2/7 = 5/7$. $P(M_2 \mid F_1) = 3/9$ and $P(M_2 \mid M_1) = 7/9$. The final answer is
>
> $$P(F_1 \mid M_2) = \frac{15}{22}.$$

Monty Hall Dilemma

Monty Hall is a game show host who shows the contestant three doors and says, "There is a prize behind one of these doors. If you choose that door, you win the prize. Please pick a door."

The contestant chooses a door (we'll assume randomly). Monty paces the stage, drawing out the torture, and just as he is about to open the chosen door, he pauses, "Hold on. Let me help you out here. I'm going to open one of the two doors you didn't choose and show you what's there."

Monty then chooses a door (we assume at random, though he will not show the prize) and opens it to reveal a donkey. He turns to the contestant again, "OK, now, do you want to stick with the door you chose originally, or do you want to switch?"

If you are the contestant, and assuming you don't want the donkey, should you switch or should you stick?

Prisoner's Dilemma

In a dark land far, far away, three prisoners are told by a sadistic king, "Tomorrow, two of you will die. One of you, who I have chosen at random, will be pardoned."

That night, one prisoner is anguishing in his cell. He calls to the warden, "Can you please tell me who will be spared?"

The warden ponders a moment, "I cannot tell you who will be spared, but I will tell you the name of one of your fellow prisoners who will not." He then names another prisoner.

Has the prisoner's chance of dying changed upon hearing this news?

3.12 Practice

The following two sections are a bit advanced. You should be able to follow them and even recreate the arguments once you've seen them. The main reason for presenting them is to demonstrate that your intuition can lead you astray when computing probabilities. They also demonstrate the use of Bayes' rule. One of the major challenges with these problems and many others is how to translate from words to mathematical notation. Again, practice makes perfect. If you take the time to understand these problems, and can recreate the arguments without looking at the solution, then you have made great progress and should be able to handle smaller problems that show up on things like exams.

3.12.1 Monty Hall's Dilemma

Restatement:
Monty Hall is a game show host who shows the contestant three doors and says, "There is a prize behind one of these doors. If you choose that door, you win the prize. Please pick a door."

The contestant chooses a door (we'll assume randomly). Monty paces the stage, drawing out the torture, and just as he is about to open the chosen door, he pauses, "Hold on. L et me help you out here. I'm going to open one of the two doors you didn't choose and show you what's there."

Monty then chooses a door (we assume at random, though he will not show the prize) and opens it to reveal a donkey. He turns to the contestant again, "OK, now, do you want to stick with the door you chose originally, or do you want to switch?"

If you are the contestant, and assuming you don't want the donkey, should you switch or should you stick?

The experiment is the complex procedure listed above, but outcomes are combinations of the location of the prize, the door the contestant chose, and the door that Monty chose. Let X be the door with the prize. Let C be the door the contestant chooses. Let M be the door Monty chooses. These are *not* events. Instead, $A = 2$, for example, is the event that the prize is behind door 2. (This notation is different from class; see below for that.)

To compute the probability that staying put gets you the prize, we need $P(C = X \mid M)$. This is an inverse probability, because it is far more natural to predict what Monty will do (M) based on C and X. Bayes' Rule

is calling. To apply Bayes' rule, we need a partition. $C = X$ and $C \neq X$ form one such partition (of many possible).

$$P(C = X \mid M) = \frac{P(M \mid C = X)P(C = X)}{P(M \mid C = X)P(C = X) + P(M \mid C \neq X)P(C \neq X)}$$

Let's compute each of these probabilities separately. First, there is only one of three ways for the contestant to guess the prize correctly $C = X$ on his own. And, if the contestant has guessed the prize, Monty has two possible choices with equal probability. If the contestant has not chosen the prize, then Monty can only open one door, because he does not want to reveal the prize.

$$\begin{aligned} P(C = X) &= \frac{1}{3} \\ P(C \neq X) &= \frac{2}{3} \\ P(M \mid C = X) &= \frac{1}{2} \\ P(M \mid C \neq X) &= 1 \end{aligned}$$

Putting all together again, we have the probability if we stick:

$$P(C = X \mid M) = \frac{\frac{1}{2}}{\frac{1}{2} + 1} = \frac{1}{3}$$

and since sticking and switching form a partition, the complement rule ($P(\overline{A}) = 1 - P(A)$) gives us the probability of winning if we switch:

$$P(C \neq X \mid M) = 1 - \frac{1}{3} = \frac{2}{3}$$

Version from class. Let A_i be the event that the prize is behind door i. Without loss of generality, suppose the contestant chooses door 1 (event B) and Monty opens door 3 (event C). Then, the probability of winning upon switching is

$$P(A_2 \mid B \cap C) = \frac{P(C \mid A_2 \cap B)P(A_2 \mid B)}{\sum_{i=1}^{3} P(C \mid A_i \mid B)P(A_i \mid B)}$$

where I have applied Bayes' rule using the conditional probability $P(\cdot \mid B)$ on both sides. Again, we consider each probability in turn.

$$P(A_i \mid B) = P(A_i)$$

by independence. The contestant's choice does not influence the location of the prize (unless something tricky is going on behind the scenes).

$$\begin{aligned} P(C \mid A_1 \cap B) &= \frac{1}{2} \\ P(C \mid A_2 \cap B) &= 1 \\ P(C \mid A_3 \cap B) &= 0 \end{aligned}$$

Putting all together, we have

$$P(A_2 \mid B \cap C) = \frac{1}{\frac{1}{2} + 1} = \frac{2}{3}$$

3.12.2 Prisoner's Paradox

Restatement:

In a dark land far, far away, three prisoners are told by a sadistic king, "Tomorrow, two of you will die, and one of you, who I have chosen at random, will be pardoned."

That night, one prisoner is anguishing in his cell. He calls to the warden, "Can you please tell me who will be spared?"

The warden ponders a moment, "I cannot tell you who will be spared, but I will tell you the name of one of your fellow prisoners who will not." He then names another prisoner.

Has the prisoner's chance of dying changed upon hearing this news?

Suppose the prisoner's are called Frank, Tom, and Jerry, and it is Frank who asks the warden for information. Let F_d be the event that Frank will die, T_d that Tom will die and J_d that Jerry will die. Let F_n, T_n, and J_n be similar events indicating that the respective prisoner is named by the warden in response to Frank's query.

First, we assume the warden names the other prisoner randomly. When Frank is doomed, the warden is forced to name the other doomed prisoner. When Frank is to be saved, the warden chooses one other prisoner to name with equal probability. Without loss of generality (WOLOG), assume the warden names Tom, then we seek $P(F_d \mid T_n)$.

At the time of the King's declaration, $P(F_d) = 2/3$. After the warden names Tom (event T_n), the prisoner knows that either he or the unnamed prisoner will live. It seems reasonable to suspect that his probability of death has been updated to $P(F_d \mid T_n) = \frac{1}{2}$, but this intuition is wrong.

Think about this problem a little while, and you'll soon see that it would really help to compute probabilities if we knew the King's decision. The King has three choices, and they form a partition $\{T_d \cap F_d, T_d \cap J_d, F_d \cap J_d\}$ with $\binom{3}{2}$, the number of ways to select two prisoners to die, elements. Then, $F_d = (T_d \cap F_d) \cup (F_d \cap J_d)$, and

$$P(F_d \mid T_n) \quad = \quad P(T_d \cap F_d \mid T_n) + P(F_d \cap J_d \mid T_n)$$

by the addition law for mutually exclusive events. Both of the conditional probabilities on the right are inverse probabilities, so we need to apply Bayes' rule:

$$P(T_d \cap F_d \mid T_n) = \frac{P(T_n|T_d \cap F_d)P(T_d \cap F_d)}{P(T_n|T_d \cap F_d)P(T_d \cap F_d)+P(T_n|T_d \cap J_d)P(T_d \cap J_d)+P(T_n|F_d \cap J_d)P(F_d \cap J_d)}$$
$$P(F_d \cap J_d \mid T_n) = \frac{P(T_n|F_d \cap J_d)P(F_d \cap J_d)}{P(T_n|T_d \cap F_d)P(T_d \cap F_d)+P(T_n|T_d \cap J_d)P(T_d \cap J_d)+P(T_n|F_d \cap J_d)P(F_d \cap J_d)}$$

Clearly, $P(T_d \cap F_d) = P(T_d \cap J_d) = P(F_d \cap J_d) = 1/3$, so all these terms cancel. As for the conditional probabilities,

$$\begin{array}{rcll} P(T_n \mid T_d \cap F_d) & = & 1 & \text{warden can only name Tom} \\ P(T_n \mid T_d \cap J_d) & = & \frac{1}{2} & \text{warden can name either Tom or Jerry} \\ P(T_n \mid F_d \cap J_d) & = & 0 & \text{warden can only name Jerry} \end{array}$$

so

$$\begin{aligned} P(T_d \cap F_d \mid T_n) &= \frac{1}{1+\frac{1}{2}} = \frac{2}{3} \\ P(F_d \cap J_d \mid T_n) &= 0 \end{aligned}$$

Frank's fate has not changed! (If only Frank could become the unnamed prisoner, however, because *that* guy has now only $\frac{1}{3}$ chance of facing the gallows. Try it and see. This wishful thinking is equivalent to Monty Hall's dilemma.)

Let's consider another scenario. Suppose the warden always chooses to name WOLOG Tom whenever he can (and Frank *knows* it), then we must distinguish $P(F_d \mid T_n)$ and $P(F_d \mid J_n)$. For the first,

$$\begin{array}{rcll} P(T_n \mid T_d \cap F_d) & = & 1 & \text{warden can only name Tom} \\ P(T_n \mid T_d \cap J_d) & = & 1 & \text{warden chooses to name Tom} \\ P(T_n \mid F_d \cap J_d) & = & 0 & \text{warden can only name Jerry} \end{array}$$

and our prisoner's fate is now $P(F_d \mid T_n) = \frac{1}{2}$, as our intuition suggested. In the other case,

$$\begin{array}{rcll} P(J_n \mid T_d \cap F_d) & = & 0 & \text{warden can only name Tom} \\ P(J_n \mid T_d \cap J_d) & = & 0 & \text{warden chooses to name Tom} \\ P(J_n \mid F_d \cap J_d) & = & 1 & \text{warden can only name Jerry} \end{array}$$

and $P(F_d \mid J_n) = 1$. This last makes sense. If Frank knows the jailor will name Tom if he can, hearing the name Jerry should strike fear in his heart. It means the warden couldn't name Tom, which means Tom is safe and Frank is doomed. In the other case, the prisoner's risk of imminent death has decreased from $\frac{2}{3}$ to $\frac{1}{2}$ exactly because the warden was free to avoid saying the name Jerry.

3.13 Loose Ends/Review

Laws and Conditional Probabilities

I have stated a few times that all probability laws also work for conditional probabilities. All you need to do is make sure that whatever conditional probability it is, say $P(\cdot \mid B)$, that the B stays put in the condition on both sides of the equality. Here are some examples of applying the rules you have learned to conditional probabilities. In all cases, C is my condition.

Law	**Example Conditional on C**
conditional probability definition	$P(A \mid B \cap C) = \frac{P(A \cap B \mid C)}{P(B \mid C)}$
law of total probability	$P(A \mid C) = \sum_{i=1}^{n} P(A \mid B_i \cap C) P(B_i \mid C)$
addition law	$P(A \cup B \mid C) = P(A \mid C) + P(B \mid C) - P(A \cap B \mid C)$
multiplication law	$P(A \cap B \mid C) = P(A \cap C) P(B \mid A \cap C)$

Independence of Complements

Corollary 20. *If A and B are independent, then $\overline{A}$ and $\overline{B}$ are independent. Likewise, $\overline{A}$ and B, A and $\overline{B}$ are independent.*

Proof.

$$\begin{array}{rcll} P(\overline{A} \cap \overline{B}) & = & P(\overline{A \cup B}) & \text{DeMorgan's Law} \\ & = & 1 - P(A \cup B) & \text{complement} \\ & = & 1 - P(A) - P(B) + P(A \cap B) & \text{addition law} \\ & = & 1 - P(A) - P(B) + P(A)P(B) & \text{multiplication law with independence} \\ & = & [1 - P(A)]\,[1 - P(B)] & \text{algebra} \\ & = & P(\overline{A})P(\overline{B}) & \text{complement} \end{array}$$

And that is one of the conditions equivalent to independence, so we are done. □

HW4 #16

I worked through HW4 #16 to demonstrate an easier use of Bayes' rule. You can see the solution there, and many like it for practice.

Test-Taking Advice

- I hope I have convinced you that your intuition can be flawed. So don't use it. Be formal, follow rules, use procedures, go step-by-step to get your answers.
- Mutually exclusive events are *not* independent events. For some reason, these are commonly confused. In fact, mutually exclusive events are definitely *dependent*, not independent. Mutually exclusive is a

property relevant to the addition law. When events are mutually exclusive, the addition law reduces to a simple sum over events. Independence is a property relevant to the multiplication law. When events are independent, the multiplication law reduces to a simple product of single events (no conditions).

- This is silly, but probabilities are numbers between 0 and 1. Double-check your answers. Another useful check is to make sure that $P(S) = 1$. Often you will be working with a partition, and you should check that the sum over that partition $\sum_{i=1}^{n} P(B_i) = 1$. Right after class, an example of where that check would have been useful came up. It is worth taking the time to verify!

3.14 Lessons Learned from Exam

Importance of Proper Notation

Correct notation is important. Mathematics is a language (a universal one at that), and if we can't agree on the rules and symbols, then we cannot communicate effectively, proofs would be ambiguous, and the whole system would break down. Here is a list of common notation mistakes and their correct versions (as I interpreted them).

WRONG	CORRECT	Comment
$A + \overline{A}$	$A \cup \overline{A}$	There is no $+$ operator for sets
$S - A = \overline{A}$	$P(S) - P(A) = P(\overline{A})$	There is no $-$ operator for sets
$P(A) \cup P(\overline{A})$	$P(A) + P(\overline{A})$	$P(\cdot)$ is a number, so traditional arithmetic symbols used
$S = 1$	$P(S) = 1$	$S = 1$ is truly ambiguous

Reading Probabilities from a Table

Event		J **Junior**	S **Senior**	**Row Totals**
E	**Econ**	0	1	1
G	**Eng**	1	4	5
M	**Math/Stat**	9	5	14
	Column Totals	10	10	**Total**$= 20$

The first thing you should do when reading a table is fill in the row and column sums. It will make what follows, really easy.

Consider the experiment where you randomly select one of the Total$= 20$ individuals and classify it by its row and column event.

Probabilities of the following events or compositions of events can be computed with ease:

- **Intersection Row Event and Column Event.** Form a ratio with corresponding entry divided by Total.

$$P(S \cap M) = \frac{5}{20}$$

- **Any Row or Column Event.** Divide the corresponding row or column sum by Total.

$$P(M) = \frac{14}{20}$$

- **Union of Events.** Because each slot in the table is mutually exclusive of all others, you just sum the matching entries. Below the first sums entries in the table, the second sums row (or column) sums.

$$P(J \cap (G \cup M)) = \frac{1+9}{20} \qquad P(G \cup M) = \frac{5+14}{20}$$

- **Conditional Probabilities $P(A \mid B)$.** Identify the entry for $A \cap B$ and put it in the numerator. Identify the row or column matching the condition, and divide by that row/column sum.

$$P(S|G) = \frac{4}{5}$$

Writing Code

It is a pain to type code in the R terminal that covers several lines. Open → New Script opens a simple editor. Type your code there, then type Ctrl+A (to select it all) and Ctrl+R (to copy, paste, and execute it at the R command line). Bugs are much easier to find and correct this way.

Also, if you don't know how to code something or you think it will take a long time, layout a plan of what you want to do and how you would use the results. Only come back to program if you have time. At least you will get partial credit for correct logic.

Detecting Combinatorics Errors

- **Match experiments.** Make sure experiment matches in numerator an denominator. The numerator counts a restricted set of the outcomes that appear in the denominator, but the outcomes must look the same.

  WRONG: 3(b)

$$\frac{\binom{5}{1}\binom{22}{1}}{\binom{27}{5}}$$

  The numerator selects 1 female and 1 male for a total of 2 individuals. The denominator selects a total of 5 individuals. Different outcomes!

- **Match Ordered vs. Unordered.** Make sure if you consider order, you count ordered outcomes in the numerator and the denominator. Or if you don't consider order, you don't consider it in neither the numerator nor the denominator. You can answer a question both ways. For example, both of the following are correct answers to one part of 3(b):

  Unordered:

$$P(0 \text{ females}) = \frac{\binom{22}{5}}{\binom{27}{5}}$$

  Ordered:

$$P(0 \text{ females}) = \frac{P_5^{22}}{P_5^{27}} = \frac{22 \times 21 \times 20 \times 19 \times 18}{27 \times 26 \times 25 \times 24 \times 23}$$

- **Ordering Can be Tricky.** But counting ordered arrangements can be a pain. For example, one might try to order the outcomes in 3(a).

  WRONG: 3(a)

$$\frac{P_1^4 P_1^{10} P_3^{13}}{P_5^{27}}$$

  Indeed you have considered all the ways to order the 3 seniors,but you have not considered the ways the sophomore and junior are arranged relative to the senior. Finding the number of ways we could order 1 sophomore, 1 junior, and 3 seniors is $\binom{5}{1\ 1\ 3}$. Thus, a corrected version that considers order is:

  CORRECTED: 3(a)

$$\frac{P_1^4 P_1^{10} P_3^{13} \binom{5}{1\ 1\ 3}}{P_5^{27}}$$

  I find it much easier just to considered unordered outcomes:

  CORRECT, VERSION II: 3(a)

$$\frac{\binom{4}{1}\binom{10}{1}\binom{13}{3}}{\binom{27}{5}}$$

- mn**-Rule Requires Disjoint Sets.** When applying the mn-rule, make sure all groups (A and B in the statement of the rule) are distinct, i.e. disjoint sets.
  - For example, groups are distinct in 3(a). Students are either sophomores, juniors, or seniors; no one is in multiple groups.

$$\begin{array}{ccc} 4 \text{ sophomores} & 10 \text{ juniors} & 13 \text{ seniors} \\ \downarrow C_1^4 & \downarrow C_1^{10} & \downarrow C_3^{13} \\ 1 \text{ selected} & 1 \text{ selected} & 3 \text{ selected} \end{array}$$
$$\downarrow$$
$$\text{apply } mn\text{-rule: } C_1^4 C_1^{10} C_3^{13}$$

  - Groups are not distinct in 3(b). Males and females are also in the group of "others". The following is WRONG:

$$\begin{array}{ccc} 5 \text{ females} & 22 \text{ males} & 25 \text{ "others"} \\ \downarrow C_1^5 & \downarrow C_1^{22} & \downarrow C_3^{25} \\ 1 \text{ selected} & 1 \text{ selected} & 3 \text{ selected} \end{array}$$
$$\downarrow$$
$$\text{apply } mn\text{-rule: } C_1^5 C_1^{22} C_3^{25}$$

    The problem is the third group "others" is not distinct from the first two. There is a dependence structure between these selection events. The order (pick male, pick 25 others, pick female vs. for example, pick female, pick male, and pick 25 others) matters and is an indication that the rule doesn't work here.

Part II

Discrete Random Variables

4 Random Variable Introduction

4.1 Definition

Definition: *random variable*

> A random variable is a real-valued function that operates on simple events ω in the sample space S of some random experiment.
>
> $$X : \omega \in S \rightarrow \Re$$

Notation

- We will use capital letters from the end of the alphabet (e.g. X, Y, Z) to represent random variables.
- We often drop the function notation, so instead of $X(\omega)$, we just write X.
- Actual values in the range of the function will be denoted by matching lower case letters. For example, for a specific experimental outcome ω_1, perhaps $X(\omega_1) = x_1$.
- Range of the function will be denoted as $X(S) = \{x : \exists\ \omega \text{ with } X(\omega) = x\}$. We might write, $X(S) = \{x_1, \ldots, x_m\}$.

Properties

- $X(\cdot)$ need not be a one-to-one function, i.e. $X(\omega_1) = X(\omega_2) = x$ for $\omega_1 \neq \omega_2$ is allowed.
- Every random variable defines a partition of the sample space S. In particular, if the sample space is finite $|S| < \infty$, then we can write $X(S) = \{x_1, \ldots, x_n\}$, where $n \leq |S|$ (less than occurs if $X(\cdot)$ is *not* one-to-one). Then, the partition is

$$\begin{aligned} B_1 &= \{\omega : X(\omega) = x_1\} \\ B_2 &= \{\omega : X(\omega) = x_2\} \\ &\vdots \\ B_n &= \{\omega : X(\omega) = x_n\} \end{aligned}$$

  Recall, that it is a partition implies $B_i \cap B_j = \varnothing$ and $\cup_{i=1}^{n} B_i = S$. Also, this generalizes to any discete sample space, just that the partition $B_1, B_2, \ldots$ may have countably many events.

- Because $X(\omega)$ operates on random outcomes, it is clearly random (perhaps the name gives this away too?). In fact, we could define a new random experiment where we generate a random outcome as usual and then map it using $X(\cdot)$. The new sample space is $X(S)$. One simple event in $X(S)$ is $\{X = x_i\}$, so it makes perfect sense to write $P(\{X = x_i\})$. However, we don't need to define new random experiments, because

$$P(\{X = x_i\}) = P(B_i)$$

  for an event B_i in the original sample space S. Thus, we can compute probabilities related to random X by using the usual probabilities of events $B_i \subset S$.

The main point so far is that nothing has changed.

Example:

Random experiment: Throw two dice. The sample space can be represented by the following table:

		Second Die					
		1	2	3	4	5	6
First Die	1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Random variable. Let Y be the sum of the two dice. Clearly $Y(S) = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, or we can map the above sample space to $Y(S)$ as follows:

		Second Die					
		1	2	3	4	5	6
First Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Induced partition. The partitions of the sample space S are visually seen to be the diagonals (lower left to upper right) in this table. Specifically

$$\begin{aligned}
B_2 &= \{\omega : Y(\omega) = 2\} = \{(1,1)\} \\
B_3 &= \{\omega : Y(\omega) = 3\} = \{(1,2),(2,1)\} \\
B_4 &= \{\omega : Y(\omega) = 4\} = \{(1,3),(2,2),(3,1)\} \\
B_5 &= \{\omega : Y(\omega) = 5\} = \{(1,4),(2,3),(3,2),(4,1)\} \\
B_6 &= \{\omega : Y(\omega) = 6\} = \{(1,5),(2,4),(3,3),(4,2),(5,1)\} \\
B_7 &= \{\omega : Y(\omega) = 7\} = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \\
B_8 &= \{\omega : Y(\omega) = 8\} = \{(2,6),(3,5),(4,4),(5,3),(6,2)\} \\
B_9 &= \{\omega : Y(\omega) = 9\} = \{(3,6),(4,5),(5,4),(6,3)\} \\
B_{10} &= \{\omega : Y(\omega) = 10\} = \{(4,6),(5,5),(6,4)\} \\
B_{11} &= \{\omega : Y(\omega) = 11\} = \{(5,6),(6,5)\} \\
B_{12} &= \{\omega : Y(\omega) = 12\} = \{(6,6)\}
\end{aligned}$$

It is easy to verify that these are all mutually exclusive and exhaustive on S.

Probabilities. In addition, because all outcomes are equally likely, we can obtain probabilities of the random variables easily by counting numerators and denominators:

$$\begin{array}{lll}
P(Y=2) = P(B_2) = \dfrac{1}{36} & P(Y=3) = P(B_3) = \frac{2}{36} & P(Y=4) = P(B_4) = \dfrac{3}{36} \\
P(Y=5) = P(B_5) = \dfrac{4}{36} & P(Y=6) = P(B_6) = \frac{5}{36} & P(Y=7) = P(B_7) = \dfrac{6}{36} \\
P(Y=8) = P(B_8) = \dfrac{5}{36} & P(Y=9) = P(B_9) = \frac{4}{36} & P(Y=10) = P(B_{10}) = \dfrac{3}{36} \\
P(Y=11) = P(B_{11}) = \dfrac{2}{36} & P(Y=12) = P(B_{12}) = \frac{1}{36} &
\end{array}$$

Example:

Suppose I ask everyone in the class to answer a "yes"/"no" question, and everyone answers independently and randomly (50:50 chance answer is "yes"). I then count the number of people with the correct answer ("yes") and let this random variable be X. What is

$$P(X=1) = \frac{\binom{24}{1}}{2^{24}}$$

I could ask you to repeat this for $X=0$, $X=1,\ldots,X=24$, but it isn't really feasible to write down all the probabilities for all the possible values in $X(S)$. Instead, can you fill in the blank for aribtrary x?

$$P(X=x) = \frac{?}{2^{24}}$$

4.2 Probability Mass Function

Probability Mass Function

Definition: *probability mass function*

A discrete random variable Y has a probability mass function (pmf) defined as

$$p(y) = P(Y=y)$$

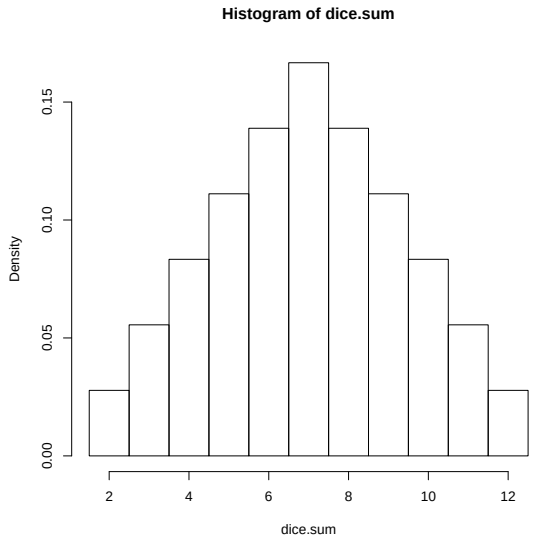
for all $y \in Y(S)$. Sometimes, when we want to emphasize *which* random variable, we'll use $p_Y(y)$ for $p(y)$.

The probability mass function is also known as the probability distribution function.

Properties

- The pmf can be displayed as a graph, table, or formula (see eq. [1] for an example of a formula). For example, a graph of the dice sum is obtained with the following `R` code:

```
> dice.sum <- matrix(data=1:6, nrow=6, ncol=6, byrow=T) +
+ matrix(data=1:6, nrow=6, ncol=6, byrow=F)
> hist(x=dice.sum, br=1:12+0.5, freq=F)
```



Aside on R

R's `matrix()` function can be used to create matrices. The first argument `data` gives the contents of the matrix, `nrow` is the number of desired rows, `ncol` is the number of desired columns, and `byrow` takes TRUE/FALSE value indicating whether the values are to be read in by row or by column. Notice, if the matrix is bigger than the number of values you provide in `data`, then it will recycle old values. For example, `matrix(0, nrow=2, ncol=2)` creates a 2×2 matrix of 0's.

R's `n:m` operator is shorthand for `seq(from=n, to=m, by=1)`. You've been using it in side `for` loops: `for(i in 1:n) { ... }`.

Notice that I have set `hist()`'s `break` argument to exactly center each of the possible values in $Y(S)$.

A table version of this pmf is

Value y	$p(y)$	Value y	$p(y)$
2	$\frac{1}{36}$	8	$\frac{5}{36}$
3	$\frac{2}{36}$	9	$\frac{4}{36}$
4	$\frac{3}{36}$	10	$\frac{3}{36}$
5	$\frac{4}{36}$	11	$\frac{2}{36}$
6	$\frac{5}{36}$	12	$\frac{1}{36}$
7	$\frac{6}{36}$		

- For values of the random variable that cannot occur, i.e. $y \notin Y(S)$, we assume $p(y) = 0$.
- The pmf fully defines the random variable. It encodes everything we understand (and don't understand because of its randomness) about the random variable. For example, if someone asks you to predict

how many students out of 24 will answer a binary question correctly if they guess, you can look at its pmf

$$p(x) = \frac{\binom{24}{x}}{2^{24}} \tag{1}$$

and make your best guess. Perhaps you would like to say you predict $X = 12$ because $p(x)$ is maximized at $p(24)$. Or you might say that you can't say for sure, but most likely 12, then maybe 11 or 13, etc.

Theorem 21. *For any discrete r.v. Y with pmf $p(y)$, the following must be true*

1. $0 \leq p(y) \leq 1$ *for all* y.
2. $\sum_{y \in Y(S)} p(y) = 1$.

Proof. Item 1 is true because $p(y) = P(Y = y)$ is a probability which satisfies the probability axioms.

Item 2 is true because Y induces a partition $\{y \in Y(S)\} = \cup_{i=1}^{\infty} B_i$. Therefore,

$$\begin{array}{rcll} \sum_{y \in Y(S)} p(y) & = & \sum_{y \in Y(S)} P(Y = y) & \text{definition of pmf} \\ & = & \sum_{i=1}^{\infty} P(B_i) & \text{induced partition} \\ & = & P(\cup_{i=1}^{\infty} B_i) & \text{partition events are mutually exclusive} \\ & = & P(S) = 1 & \text{partition is exhaustive} \end{array}$$

by the axioms and definition of partition. □

One of the Most Important Modern Ideas

Suppose you use R to throw a pair of dice n times and for each outcome you compute the sum of the showing faces. Let $X_1, \ldots, X_n$ be the random variables, and R generates (independent) realizations $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

We can summarizes the n trials as 11 numbers $n_2, n_3, \ldots, n_{12}$, where n_i is the number of times you observe i in the sequence $X_1, \ldots, X_n$.

Recall the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Using the summary data n_i, we can rewrite it as

$$\bar{X} \quad = \quad \frac{1}{n} \sum_{i=2}^{12} i n_i = \sum_{i=2}^{12} i \frac{n_i}{n}$$

But you already know that as the sample size increases, $\frac{n_i}{n} \to P(X = i)$. (Remember flipping a coin. If we flipped a coin 1000 times and recorded the number of heads, the proportion formed by dividing the number of heads by 1000 came closer and closer to 0.5, which is exactly $P(\text{head})$.) Thus,

$$\bar{X} \to \sum_{i=2}^{12} i P(X = i)$$

the sample mean approaches the quantity on the right as the number of repetitions increases. The quantity on the right is the expected value of X, which we will define next time. If the pmf you have assumed for your random variable is correct, then the expected value is the population mean. Thus, the sample mean gets closer and closer to that elusive population mean that I introduced you to at the beginning of class, the one I told you we can never know. Well, we can never know it, but we can get darned close just by generating $X_1, X_2, \ldots, X_n$! These random variables can come from (1) simulation in R, (2) sampling from the population, (3) emulation in a laboratory, or (4) your next creative idea?

More importantly, for the bigger picture, in deriving this result, you can see we were able to estimate/approximate the pmf $\frac{n_i}{n} \approx P(X = i)$. Since the pmf encodes all information about the process, we now have an estimate of everything there is to know about the entire process. That is powerful.

Why is this concept so important? Suppose you are studying a system so complex, so intricate that you have no clue what the pmf of an interesting quantity X you can measure on that system. Sometimes, you may not even know the sample space S. Examples: (1) explain the differences between the human and neanderthal genomes, let X be the number of differences required to explain human speech, (2) predict how many days before the recession is over, let X be the number of days, (3) let X be the number of dollars Iowa collects in taxes in 2009, (4) let Y be the number of cars a bridge can tolerate before it collapses, (5) predict the amount of CO_2 Z in the atmospher in 2050.

In each case, you can't collect directly relevant data to answer the question, or if you can you can't collect replicates of that data. On the other hand, you may have a computer model or mathematical model that incorporates everything that is so far known about the process. You can run this model and see what it says about the number of (1) differences, (2) months, (3) dollars, (4) cars, or (5) CO_2 molecules.

If there is a lot of uncertainty in the model then no two runs of the simulator will give the same result, so running it once is hardly useful. However, if you run it again-and-again, say n times, then you can estimate the pmf and make predictions about the quantity of interest. You can predict the number of (1) mutations it takes to explain speech, (2) months the recession will last, (3) dollars Iowa will collect and therefore can spend, (4) the number of cars you can let on the bridge and keep everyone safe, and (5) and how dire the CO_2 situation will be in 2050.

Very relevant, very powerful.

4.3 Expected Value

Definition: *expected value*

> If random variable Y has probability mass function $p_Y(y)$, then the *expected value* (or *expectation*) of Y is
>
> $$E[Y] = \sum_{y \in Y(S)} y p_Y(y)$$

Remark:

- **Convergence.** Technically, because the list of possible values Y (i.e. $Y(S)$) may be infinite, we need to worry about whether the above sum converges. If it does not, i.e. $\sum_{y \in Y(S)} y p_Y(y) = \infty$, then $E[Y]$ is not defined. In technical definitions, $E[Y]$ is defined as long as $\sum_y y p_Y(y)$ is absolutely convergent, i.e. $\sum_y |y| p_Y(y) < \infty$.
- **Relation to Sample Mean $\bar{X}$.** Recall the sample mean for any sample of data $X_1, \ldots, X_n$ collected from a population is
  $$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$
  Last time, we argued that $\bar{X} \to E[X]$ as sample size n increases. This is true as long as X_i are random variables that match X. In other words, the assumptions we make in deriving the pmf $p_X(x)$ must be satisfied when obtaining X_i. For example, when flipping a coin, we assume that the coin is fair. If our assumptions are wrong then there is no telling where $\bar{X} \to$ converges.
- **Relation to Population Mean.** Recall when we mentioned sample mean $\bar{X}$, we also discussed the population mean μ as the unknowable mean of the *entire* population. *When all assumptions are met*, then not only should $\bar{X}$ be close to $E[X]$, but $E[X] = \mu$ *is* the population mean.

- **Modeling Reality.** In real life, no model of reality is exactly correct. Even the assumption of a perfectly fair coin cannot be satisfied by any coin (they are all flawed somehow), so though we can compute $E[X]$, we can still never know the population mean μ. Nevertheless, if the model is very close to correct, then $E[X] \approx \mu$ is very close to the population mean. Further, if we sample enough $X_1, X_2, \ldots$, then $\bar{X} \to \mu$ can also get arbitrarily close to the population mean. Often we neglect the difference between reality and the model, in which case we assume $E[X] = \mu$. In this case $\bar{X} \to E[X] = \mu$.

Approximation	True When:
$\bar{X} \approx \mu$	always if n is large enough
$E[X] = \mu$	only if model and its assumptions are correct
$\bar{X} \approx E[X]$	only if model and its assumptions are correct and n is large enough

Functions of Random Variables

Let X_1 be the number of die 1, X_2 the number of on die 2, then the sum of two dice is

$$X = X_1 + X_2$$

X is a function of two random variables. We might define other functions, e.g. $X = \max(X_1, X_2)$.

In general, we can take and function $g(\cdot)$ and use it to define a new discrete random variable

$$Y = g(X)$$

for any discrete random variable X. In fact, Y is a composite function that operates on $\omega \in S$, e.g. as a function $Y(\omega) = g[X(\omega)] = (g \circ X)(\omega)$.

We already know that X induces a partition $B_1, B_2, \ldots$ on the sample space of the random experiment S, but it is also true that Y induces a partition $C_1, C_2, \ldots$ on the sample space.

Example:

Let X be the sum of two dice. The sample space is $S = \{(1,1),(1,2),(2,1),\ldots\}$. The range of X is $X(S) = \{2, 3, \ldots, 12\}$. X induces the partition

$$\begin{aligned} B_2 &= \{\omega : X(\omega) = 2\} = \{(1,1)\} \\ &\vdots \\ B_{12} = \{\omega : X(\omega) = 12\} &= \{(6,6)\} \end{aligned}$$

Now, introduce $g(x) = x\%3$, the modulus operator that returns the remainder after dividing by the second argument, and let $Y = g(X)$ be the random variable obtaine dy applying $g(\cdot)$ to X. For example, $g(2) = 2, g(3) = 0, g(4) = 1, \ldots, g(12) = 0$.

Clearly, the range of Y is $Y(S) = Y[X(S)] = \{0, 1, 2\}$. Further, Y induces a partition

$$\begin{aligned} C_0 &= \{\omega : Y[X(\omega)] = 0\} = \{(1,2),(2,1),(1,5),(5,1),\ldots\} \\ C_1 &= \{\omega : Y[X(\omega)] = 1\} = \{(1,3),(3,1),(2,2),(1,6),\ldots\} \\ C_2 &= \{\omega : Y[X(\omega)] = 2\} = \{(1,1),(1,4),(4,1),(2,3),\ldots\} \end{aligned}$$

We also notice that every event in the partition C_0, C_1, C_2 can be written as a union of events in the partition $B_1, B_2, \ldots$. For example,

$$C_0 = \{\omega : X(\omega) = 3 \text{ or } X(\omega) = 6 \text{ or } X(\omega) = 9 \text{ or } X(\omega) = 12\} = B_3 \cup B_6 \cup B_9 \cup B_{12}$$

Theorem 22. *For a discrete r.v. Y with pmf $p_Y(y)$ and function $g(y)$, then*

$$E[g(Y)] = \sum_{y \in Y(w)} g(y) p_Y(y)$$

Proof. Let B_y be the partition induced by Y and C_x be the partition induced by X.

$$\begin{array}{rcll} E[X] & = & \sum_x x p_X(x) & \text{definition of } E[X] \\ & = & \sum_x x P(C_x) & \text{definition of pmf of } X \\ & = & \sum_x x P(\cup_{y:g(y)=x} B_y) & C_x \text{ is a union of } B_y \\ & = & \sum_x x \sum_{y:g(y)=x} p_Y(y) & \text{definition of pmf of } Y \\ & = & \sum_x \sum_{y:g(y)=x} x p_Y(y) & \text{move } x \text{ inside second sum} \\ & = & \sum_x \sum_{y:g(y)=x} g(y) p_Y(y) & \text{g(y)=x} \\ & = & \sum_y g(y) p_Y(y) & \text{double sum is sum over all } y \in Y(S) \end{array}$$

□

Definition: *variance*

If Y is discrete random variable with expectation $E[Y] = \mu$, its variance is

$$V(Y) = E\left[(Y - \mu)^2\right]$$

with standard deviation $+\sqrt{V(Y)}$.

Example:

1. Suppose you have a finite r.v. Y with the following pmf (displayed in table format)

y	$p_Y(y)$
0	0.2
2	0.3
18	0.1
20	0.4

   Compute the expectation $E[Y]$ and variance (Y).

$$\begin{array}{rcl} E[Y] & = & 0 \times 0.2 + 2 \times 0.3 + 18 \times 0.1 + 20 \times 0.4 \\ & = & 10.4 \\ V(Y) & = & (0 - 10.4)^2 \times 0.2 + (2 - 10.4)^2 \times 0.3 + (18 - 10.4)^2 \times 0.1 + (20 - 10.4)^2 \times 0.4 \\ & = & 85.44 \end{array}$$

   The standard deviation is $\sqrt{85.44} = 9.24$.

2. Compute the expectation of $X = g(Y) = Y\%3$, where Y is the sum of two dice. We can use the theorem.

$$
\begin{aligned}
E[X] &= E[g(Y)] \\
&= \sum_y g(y)p_Y(y) \\
&= g(2) \times \frac{1}{36} + g(3) \times \frac{2}{36} + g(4) \times \frac{3}{36} + g(5) \times \frac{4}{36} \\
&= + g(6) \times \frac{5}{36} + g(7) \times \frac{6}{36} + g(8) \times \frac{5}{36} + g(9) \times \frac{4}{36} \\
&= + g(10) \times \frac{3}{36} + g(11) \times \frac{2}{36} + g(12) \times \frac{1}{36} \\
&= 2 \times \frac{1}{36} + 0 \times \frac{2}{36} + 1 \times \frac{3}{36} + 2 \times \frac{4}{36} \\
&= + 0 \times \frac{5}{36} + 1 \times \frac{6}{36} + 2 \times \frac{5}{36} + 0 \times \frac{4}{36} \\
&= + 1 \times \frac{3}{36} + 2 \times \frac{2}{36} + 0 \times \frac{1}{36} \\
&= 1
\end{aligned}
$$

Properties of Expectation

Theorem 23. *For discrete r.v. Y with pmf $p_Y(y)$ and c constant, then $E[c] = c$.*

Proof.

$$
\begin{aligned}
E[c] &= \textstyle\sum_y cp_Y(y) && \text{definition of expectation} \\
&= c\textstyle\sum_y p_Y(y) && c \text{ is constant} \\
&= c\textstyle\sum_y P(B_y) && \text{definition of pmf} \\
&= c \times 1 && \textstyle\sum_y P(B_y) = 1 \text{ because } B_y \text{ is a partition, i.e. exhaustive} \\
&= c
\end{aligned}
$$

□

Theorem 24. *For discrete r.v. Y with pmf $p_Y(y)$, c constant, then $E[cY] = cE[Y]$.*

Proof.

$$
\begin{aligned}
E[cY] &= \textstyle\sum_y cyp_Y(y) && \text{definition of expectation and theorem for } E[g(Y)] \\
&= c\textstyle\sum_y yp_Y(y) && c \text{ is constant} \\
&= cE[Y] && \text{definition of expectation}
\end{aligned}
$$

□

Theorem 25. *For discrete r.v. Y with pmf $p_Y(y)$, $a_1, \ldots, a_n$ constants and $g_1(y), \ldots, g_n(y)$ arbitrary functions, then*

$$
E\left[\sum_{i=1}^n a_i g_i(y)\right] = \sum_{i=1}^n a_i E[g_i(Y)]
$$

Proof.

$$
\begin{aligned}
E\left[\textstyle\sum_{i=1}^n a_i g_i(y)\right] &= \textstyle\sum_y \sum_{i=1}^n a_i g_i(y) p_Y(y) && \text{theorem for } E[g(Y)] \\
&= \textstyle\sum_{i=1}^n \sum_y a_i g_i(y) p_Y(y) && \text{exchange sums, ok if converge} \\
&= \textstyle\sum_{i=1}^n a_i \sum_y g_i(y) p_Y(y) && a_i \text{ is constant wrt } y \\
&= \textstyle\sum_{i=1}^n a_i E[g_i(Y)] && \text{theorem for } E[g(Y)]
\end{aligned}
$$

□

Main conclusion. $E[\cdot]$ is a linear operator.

Theorem 26. *If Y is a discrete random variable with pmf $p_Y(y)$ and $E[Y] = \mu$, then*

$$V(Y) = E[Y^2] - \mu^2$$

Proof.

$$\begin{array}{rcll} V(Y) & = & E\left[(Y-\mu)^2\right] & \text{definition of variance} \\ & = & E\left[Y^2 - 2\mu Y + \mu^2\right] & \text{FOIL} \\ & = & E\left[Y^2\right] - 2\mu E\left[Y\right] + \mu^2 & \text{linearity of expectation} \\ & = & E\left[Y^2\right] - 2\mu^2 + \mu^2 & \text{definition of } E[Y] \end{array}$$

□

Theory Exercise

Example:

For a random variable Y and constant c, what is $V(cY)$?

$$\begin{array}{rcll} V(cY) & = & E\left[(cY - E[cY])^2\right] & \text{definition of variance} \\ & = & E\left[(cY - cE[Y])^2\right] & \text{linearity of expectation} \\ & = & E\left[c^2(Y - E[Y])^2\right] & \text{algebra} \\ & = & c^2 E\left[(Y - E[Y])^2\right] & \text{linearity of expectation} \\ & = & c^2 V(Y) & \text{definition of variance} \end{array}$$

Example:

1. Two different sizes of shipping containers are available. A percent 30% of containers are size $8 \times 10 \times 30$ and the rest are size $8 \times 10 \times 40$. Assuming the container is filled to capacity, what is the expected volume per container shipped?
   Let I be the event that a container of the first size is selected, II for the other type. The random experiment is to select a container and use it to ship a load. The sample space is $S = \{I, II\}$. Let X be the length of the container selected. Let Y be the volume shipped, then clearly
   $$Y = 8 \times 10 \times X = 80X$$
   and we see $Y = g(X)$ is a function of X. We seek $E[Y]$.
   Expectation by Definition. By properties of expectation
   $$E[Y] \quad = \quad E[g(X)] = \sum_x \in X(S) g(x) p_X(x)$$
   The range of X is $X(S) = \{30, 40\}$. The pmf of X is
   $$g_X(30) = 0.3 \qquad\qquad g_X(40) = 0.7$$
   yielding
   $$E[Y] = g(30)g_X(30) + g(40)g_X(40) = 80 \times 30 \times 0.3 + 80 \times 40 \times 0.7 = 2960.$$
   Expectation by Properties. A quicker solution, recognizes that $g(X)$ is a linear function, so
   $$E[Y] = E[g(X)] = E[80X] = 80E[X]$$
   and
   $$E[X] = 30 \times 0.3 + 40 \times 0.7 = 37$$

so $E[Y] = 80 \times 37 = 2960$

Variance by Definition Let's also compute $V(Y)$.

$$V(Y) = (80 \times 30 - 2960)^2 \times 0.3 + (80 \times 40 - 2960)^2 \times 0.7 = 134400$$

Variance by Properties But $V(Y)$ can also be computed with our most recent result.

$$V(Y) = V[g(X)] = V[80X] = 80^2 V(X)$$

with $V(X) = E[X^2] - (E[X])^2 = 30^2 \times 0.3 + 40^2 \times 0.7 - 37^2 = 21$, so $V(Y) = 80^2 * 21 = 134400$.

2. A customer wants to buy a $\$40,000$ insurance policy, and you work for the insurance company. How much should you charge if you want to break even *on average.* You will make your choice based on the following information on historical claims: a fraction 0.001 of policies had 100% claims filed against them, another fraction 0.01 of policies had 50% claims filed against them, all remaining policies had no claims filed.

   The random experiment is to sell and insurance policy and record gain or loss after coverage expires. The sample space is $S = \{A, B, D\}$, where A is the event that a 100% claim is filed, B is the event that a 50% claim is filed, and D is the event that no claim is filed.

   Let Y be the fraction loss claim filed against the policy. We assume $Y(S) = \{0, 0.5, 1\}$. Let X be the amount earned on the policy by your company. Clearly,

$$X = c - 40000Y$$

   where c is the price you charge for the policy. We seek to set $E[X] = 0$ and use the resulting equation to solve for c.

$$\begin{array}{rcll} E[X] & = & E[c - 40000Y] & \text{as per problem statement} \\ & = & c - 40000E[Y] & \text{linearity of expectation} \end{array}$$

   To compute $E[Y]$, we'll need the pmf $p_Y(y)$.

$$p_Y(0.5) = 0.01 \qquad p_Y(1) = 0.001 \qquad p_Y(0) = 1 - 0.01 - 0.001$$

   Then,

$$\begin{array}{rcll} E[Y] & = & 0.5p_Y(0.5) + 1p_Y(1) + 0p_Y(0) & \text{definition of expectation} \\ & = & 0.5 \times 0.01 + 1 \times 0.001 + 0 \times \text{whatever} & \\ & = & 0.006 & \\ E[X] & = & c - 40000 \times 0.006 & \\ c & = & 240 & \end{array}$$

3. Suppose $Y = X + 1$. How is $E[Y]$ related to $E[X]$? Obviously, by linearity, $E[Y] = E[X + 1] = E[X] + 1$, so $E[Y] > E[X]$. How is $V(Y)$ related to $V(X)$?

$$\begin{array}{rcl} V(Y) & = & V(X + 1) \\ & = & E\left[(X + 1 - E[X + 1])^2\right] \\ & = & E\left[(X + 1 - E[X] - 1])^2\right] \\ & = & E\left[(X - E[X]])^2\right] \\ & = & V(X) \end{array}$$

   The above example suffices to prove the following lemma (replace 1 with any constant c)

   Lemma 27. *For random variable X and constant c, $V(X + c) = V(X)$.*

4. Suppose $Y = \ln X$. Is it true that $E[Y] = \ln(E[X])$? No.

5 Bernoulli Random Variable

We now turn our attention to random variables with names. These are random variables that show up over and over again in applications. Each of these random variables will be characterized by, at least, its pmf as well as expectation and variance.

Definition: *Bernoulli Random Variable*

> Y is said to be a Bernoulli random variable if and only if it has the following pmf
>
> $$p_Y(1) = p \qquad\qquad p_Y(0) = 1 - p$$
>
> in which case we say $Y \sim$ Bernoulli(p).

Typically, we refer to $q = 1-p$, but there is only one parameter p used to define the Bernoulli random variable. From it's pmf, we can see a Bernoulli random variable has only two possible outcomes $Y(S) = \{0, 1\}$.

Here are some examples of situations that are well-modeled by the Bernoulli random variable. Notice, in particular, that we can take any random experiment with only two simple events ω_1 and ω_2 and define $Y(\omega_1) = 1$ and $Y(\omega_2) = 0$ to get a Bernoulli random variable.

1. Flip a coin and record if it is heads ($Y = 1$) or not ($Y = 0$).
2. Poll a random inviduals and record whether they support ($Y = 1$) a candidates or ($Y = 0$) do not.
3. Test a product whether it is defective ($Y = 0$) or not ($Y = 1$).
4. Come to class ($Y = 1$) or not ($Y = 1$). Seems like $p \approx 0.5$ for Stat 341.

Expectation

The expectation is

$$E[Y] \quad = \quad 1 * p + 0 * (1 - p) = p$$

Variance

The variance is

$$\begin{aligned} V(Y) &= (1-p)^2 p + (0-p)^2(1-p) \\ &= p - 2p^2 + p^3 + p^2 - p^3 \\ &= p - p^2 = p(1-p) = pq \end{aligned}$$

6 Binomial Random Variable

6.1 Probability Mass Function

Definition: *Binomial random experiment*

> A *binomial random experiment* is an experiment where the following properties are satisfied.
>
> 1. There are a fixed n identical trials.
> 2. Each trial results in a success 1 or failure 0.
> 3. The probability of success for all trials is p. Let $q = 1 - p$.
> 4. Trials are independent.

5. The random variable of interest is the number of successes.

It is often taught that a binomial random variable is the number of successes in n independent and identically distributed (iid) Bernoulli trials with probability of success p. *Identically distributed* implies that the probability of success p must be constant across trials.

The sample space of a Binomial random experiment is $S = \{\overbrace{0\cdots\cdots0}^{n \text{ times}}, 1\overbrace{0\cdots\cdots0}^{n-1 \text{ times}}, 01\overbrace{0\cdots\cdots0}^{n-2 \text{ times}}, \ldots, \overbrace{1\cdots\cdots1}^{n \text{ times}}\}$.
The range of the random variable $Y(S) = \{0, 1, \ldots, n\}$.

Outcome	**Probability**	Y	**Count**
$\overbrace{0\cdots\cdots0}^{n \text{ times}}$	$(1-p)^n$	0	1
$1\overbrace{0\cdots\cdots0}^{n-1 \text{ times}}$	$p(1-p)^{n-1}$	1	$\binom{n}{1}$
$01\overbrace{0\cdots\cdots0}^{n-2 \text{ times}}$	$p(1-p)^{n-1}$	1	
$\vdots$	$\vdots$	$\vdots$	
$\overbrace{0\cdots\cdots01}^{n-1 \text{ times}}$	$p(1-p)^{n-1}$	1	
$11\overbrace{0\cdots\cdots0}^{n-2 \text{ times}}$	$p^2(1-p)^{n-2}$	2	$\binom{n}{2}$
$101\overbrace{0\cdots\cdots0}^{n-3 \text{ times}}$	$p^2(1-p)^{n-2}$	2	
$\vdots$	$\vdots$	$\vdots$	
$\overbrace{0\cdots\cdots011}^{n-2 \text{ times}}$	$p^2(1-p)^{n-2}$	2	
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\overbrace{1\cdots\cdots1}^{n \text{ times}}$	p^n	n	1

Putting all this together, we derive the probability mass function for Y

$$P(Y = y) = p_Y(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

Definition: *binomial random variable*

> A discrete random variable Y is called a *binomial random variable* with n trials and probability of success p if its pmf is
>
> $$p_Y(y) = \binom{n}{y} p^y (1-p)^{n-y}$$
>
> We say $Y \sim \text{Binomial}(n, p)$.

6.2 Binomial in R

Binomial Distribution in R

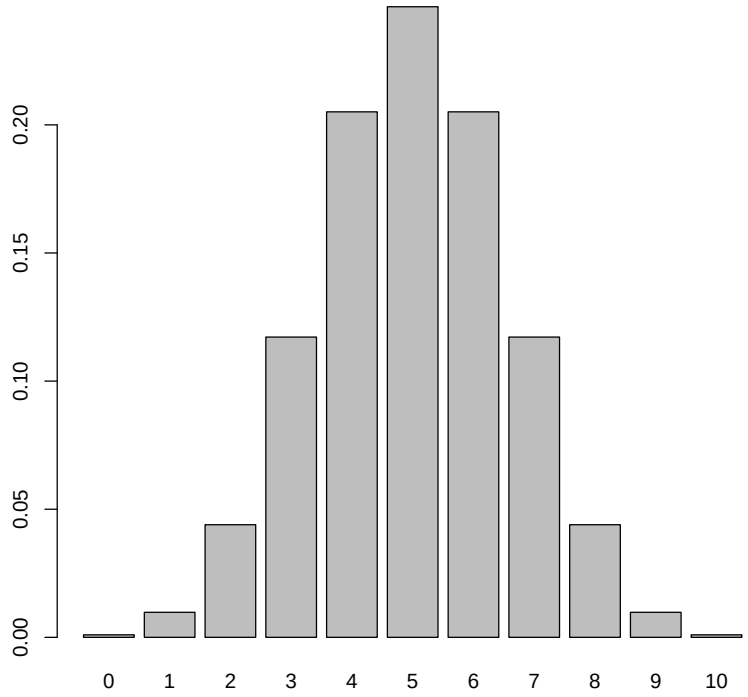
The following functions are available in R for working with the binomial distribution. In the following, it is assumed $Y \sim \text{Binomial}(n, p)$.

Function Name	What it Computes
`dbinom(x, size=n, prob=p)`	$P(Y = x)$
`pbinom(q, size=n, prob=p)`	$P(Y \leq x)$
`qbinom(p, size=n, prob=p)`	Not discussed
`rbinom(n, size=m, prob=p)`	Generate n $Y \sim$ Binomial(m, p)

For the `rbinom()` function, I had to change the number of trials to avoid confusion with its first argument, known as n.

Computing Probabilities

The following is a plot of the pmf for $Y \sim$ Binomial$(10, 0.5)$. We notice it is symmetric, but unless $p = 0.5$, binomial distributions are not symmetric.



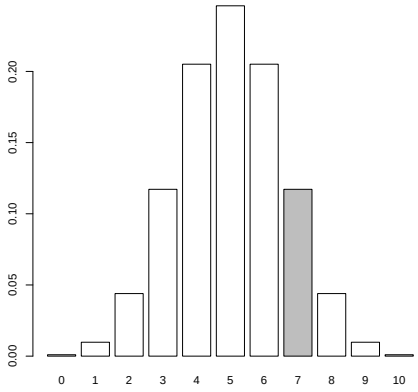
The next three plots shade the area computed by the command that precedes them.

6.3 Expectation/Variance

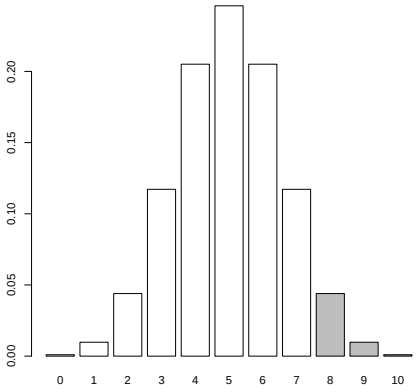
Properties of Binomial Random Variable

Theorem 28. *If $Y \sim$ Binomial(n, p), then*

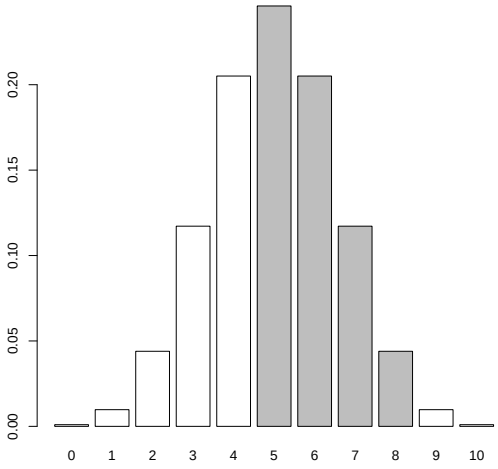
$$\begin{aligned} E[Y] &= np \\ V(Y) &= np(1-p) \end{aligned}$$



(a) `dbinom(7, size=10, prob=0.5)`



(b) `1-dbinom(8, size=10, prob=0.5)`



(c) `dbinom(8, size=10, prob=0.5) - dbinom(4, size=10, prob=0.5)`

Proof. The proof illustrates an important trick that is used frequently to simplify expressions for expectations and variances. This trick is also used to find expressions for some mathematical series.

$$\begin{array}{rcll} E[Y] & = & \sum_y y p_Y(y) & \text{definition of expectation} \\ & = & \sum_{y=0}^n y\binom{n}{y}p^y(1-p)^{n-y} & \text{pmf for binomial rv} \\ & = & \sum_{y=1}^n y\binom{n}{y}p^y(1-p)^{n-y} & y=0 \text{ term i } 0 \\ & = & \sum_{y=1}^n y\frac{n!}{y!(n-y)!}p^y(1-p)^{n-y} & \text{definition of combination} \\ & = & np\sum_{y=1}^n \frac{(n-1)!}{(y-1)!(n-y)!}p^{y-1}(1-p)^{n-y} & \text{rearrangement (trick in mind)} \\ & = & np\sum_{x=0}^{n-1} \frac{(n-1)!}{x!(n-x-1)!}p^{x}(1-p)^{n-x-1} & \text{change of variable } x=y-1 \\ & = & np\sum_{x=0}^{n-1} p_X(x) & X \sim \text{Binomial}(n-1,p) \\ & = & np & \text{sum} = 1 \text{ by LTP} \end{array}$$

□

To prove the variance, we need another result:

Lemma 29. *For any random variable Y,*

$$E[Y(Y-1)] = E[Y^2] - E[Y] = V(Y) + (E[Y])^2 - E[Y]$$

Proof. By the rules for expectations. □

With this in hand, see if you can find $E[Y(Y-1)]$ for $Y \sim$ Binomial using the same trick as above. The book shows the details, so you can check yourself.

6.4 Examples

1. [Numbers have changed from class.] Suppose an oil company has enough money to finance 30 explorations, but the probability of a successful oil exploration is only 0.01. Assume explorations are independent, and find the mean and variance of the number of successful explorations.

   Solution.

   The first question to ask yourself: Is this a binomial experiment?

   If we take "exploration" to be a "trial," the answer is yes. The number of explorations $n = 30$ is given up front. Explorations either end in "success" or "failure." The probability $p = 0.01$ is constant for all explorations. We are told the explorations are independent. Finally, we are interested in Y the number of successful explorations.

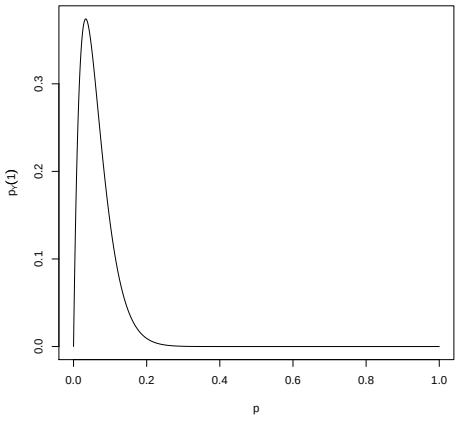
   Thus, $Y \sim$ Binomial$(30, 0.01)$ and by the last theorem,

   $$E[Y] = np = 30 \times 0.01 = 0.3 \qquad V(Y) = np(1-p) = 30 \times 0.01 \times 0.99 = 0.297$$

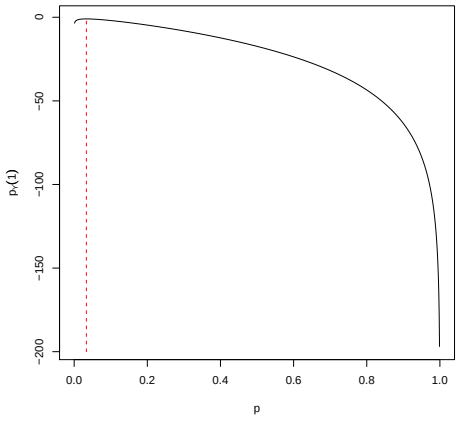
   It is quite evident from these numbers that there is a fairly good chance that the company will fail to find any oil after 30 explorations. The empirical rule suggests that we expect Y to fall in the range $(E[Y] - 2\sqrt{V(Y)}, E[Y] + 2\sqrt{V(Y)}) = (-0.7899541, 1.389954)$ with 95% probability. Since Y is actually an integer in $Y(S) = \{0, 1, 2, \ldots, 30\}$, these results suggest that 95% of the time, we expect $Y = 0$ or $Y = 1$. The company's chances do not look great. We might worry that the empirical rule may not be well-satisifed by the Binomial distribution, but if we instead compute

   $$p_Y(0) + p_Y(1) = 0.7397004 + 0.2241516 = 0.963852$$

   using `dbinom(0, size=30, prob=0.01) + dbinom(1, size=30, prob=0.01)` or, more efficiencly, `pbinom(1, size=30, prob=0.01)`, our worry about the appropriateness of the *empirical rule* is relieved.



(d) likelihood function



(e) log likelihood function

Figure 1: Binomial likelihood and log likelihood function for $n = 30$ and observed data $y_0 = 1$

2. To spice things up a bit, let's compute the cost of the explorations. Suppose the company must pay \$10,000 to prepare the exploration equipment. Then, for every successful exploration, they pay \$20,000, and for every failed exploration, they pay \$15,000. What is the expected total cost of exploration.

   Solution.

   Let X be the total cost of exploration. X is a random variable because it depends on the random outcome of the exploration experiment. It clearly matters how many explorations are successful, and how many are not. In fact, if we can write X as some function of Y (the number of succesful explorations), then we can use results for expectation to compute the mean cost.

   It should not be hard to verify that if there are Y successes, then there must have been $30 - Y$ failures. With this information, we know

   $$X = 10000 + 20000 \times Y + 15000 \times (30 - Y)$$

   so

   $$E[X] = E[10000 + 20000 \times Y + 15000 \times (30 - Y)] = 460000 + 5000E[Y] = 461500$$

Lesson. In the general binomial experiment, if we observe Y successes, then there must have been $n - Y$ failures. If successes "cost" c_1 and failures "cost" c_2, then the total "cost" is

$$X = Yc_1 + (n - Y)c_2$$

I put "cost" in quotes, because I hope you can see how this would generalize to any kind of measure on failures/successes.

6.5 Maximum Likelihood Estimation

Problem. Suppose that you know $Y \sim \text{Binomial}(n, p)$, but you do not know the probability of success p. Can you collect data to learn something about p? Of course. That is what statistics is all about.

Suppose you carry out n experiments and observe y_0 successes. An intuitive best guess of p would be

$$\hat{p} = \frac{y_0}{n}$$

the number of times a success happened over the total number of trials. We used this estimate for the probability of tails when we used R to flip coins. We will now find this estimate through mathematical arguments.

Assuming that $Y \sim \text{Binomial}(n, p)$, then the probability of observing y_0 successes is

$$P(Y = y_0) = p_Y(y_0) = \binom{n}{y_0} p^{y_0}(1-p)^{n-y_0}$$

Whereas before we used this formula to compute the probability of an outcome given some p, we now cannot obtain a number on the right-hand-side because p is unknown.

Definition: *binomial likelihood*

> The *likelihood* of the observed data $Y = y_0$ when we assume $Y \sim \text{Binomial}(n, p)$ with p unknown is
>
> $$L(p) := P(Y = y_0) = \binom{n}{y_0} p^{y_0}(1-p)^{n-y_0}$$
>
> Notice n and y_0 are known, so it is a function of the unknown parameter p.

Definition: *binomial log likelihood*

> The *log likelihood* of the observed data $Y = y_0$ when we assume $Y \sim \text{Binomial}(n, p)$ is
>
> $$l(p) := \ln P(Y = y_0) = \ln \binom{n}{y_0} + y_0 \ln p + (n - y_0) \ln(1-p)$$
>
> It is also a function of the unknown parameter p.

Consider the specific example $n = 30$ and $y_0 = 1$ (oil exploring example, but with p unknown). The likelihood function is shown in Fig. 1(d). The log likelihood function is shown in Fig. 1(e).

Lemma 30. *The p that maximizes $L(p)$ also maximizes $l(p)$.*

Proof. Suppose p_1 maximizes $L(p)$ and $p_2 \neq p_1$ maximizes $l(p_2)$. By the first claim,

$$L(p_1) \geq L(p_2)$$

But, $\ln(x)$ is a monotonically increasing function (see Fig. 118), so for any $a \geq b$, we have

$$\ln a \geq \ln b$$

(You can see this also by observing $\frac{d \ln(x)}{dx} = \frac{1}{x} > 0$.) Thus,

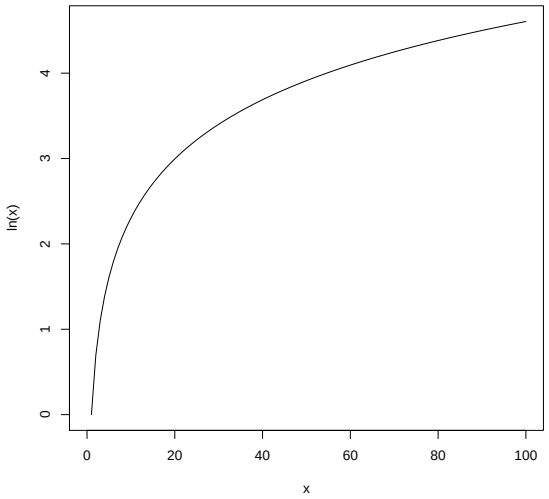
$$\ln L(p_1) \geq \ln L(p_2) = l(p_2)$$

but then $l(p_1) \geq l(p_2)$, which contradicts that p_2 maximizes $l(p)$ unless $l(p_2) = l(p_1)$. But, if $l(p_1) = l(p_2)$, then $p_1 = p_2$ because $\ln(x)$ is a *strictly* monotonic function. □

Preliminaries aside, we are now ready to estimate p. It makes sense to find the value of p that maximizes the probability of observing y_0. It makes no sense to choose an estimate of p that makes the data we ended up observing y_0 *unlikely*. Instead, we want the p that makes the actual data *more likely*. In other words, we seek the estimate $\hat{p}$ that maximizes the *likelihood*, or equivalently, the *log likelihood*.

Definition: *maximum likelihood estimate $\hat{p}$*

> The *maximum likelihood estimate* of p is the value of p that maximizes the log likelihood, i.e.
>
> $$\hat{p} = \text{argmax}_p l(p)$$



Theorem 31. *The maximum likelihood estimate for p assuming $Y \sim$ Binomial(n, p) and observing data $Y = y_0$ is*

$$\hat{p} = \frac{Y}{n}$$

Proof. Let's find $\hat{p}$ for the Binomial data. To find the maximum (or minimum) of a function, we take its derivative and set it equal to 0.

$$\begin{aligned}
l(p) &= \ln \tbinom{n}{y_0} + y_0 \ln p + (n - y_0) \ln(1 - p) && \text{definition of binomial log likelihood} \\
\tfrac{dl(p)}{dp} &= 0 && \text{calculus: finding min/maxima} \\
\tfrac{y_0}{\hat{p}} - \tfrac{n-y_0}{1-\hat{p}} &= 0 && \text{calculus: } \tfrac{d \ln x}{dx} = \tfrac{1}{x} \\
y_0(1 - \hat{p}) - (n - y_0)\hat{p} &= 0 && \text{multiply both sides by } \hat{p}(1 - \hat{p}) \\
\hat{p} &= \tfrac{y_0}{n} && \text{rearrange}
\end{aligned}$$

and so we end up with the very estimate we guessed at the beginning. To verify it is a maximum, we examine the second derivative

$$\frac{d^2 l(p)}{dp^2} = -\frac{y_0}{p^2} - \frac{n - y_0}{(1 - p)^2} \leq 0$$

which is clearly negative because $y_0, n-y_0, p, 1-p$ are all positive numbers by the restrictions that $0 \leq p \leq 1$ and $y_0 \in Y(S) = \{0, 1, \ldots, n\}$. □

For the example $n = 30$ and $y = 1$, the maximum likelihood estimate of p is

$$\hat{p} = \frac{1}{30}$$

shown as the dotted red line in Fig. 1(e).

Important concept. The idea to estimate parameters by choosing the values that maximizes the probability (or log probability) of observing the data is fundamental in statistics. Can you see how valuable it is to be able to collect data and estimate fundamental parameters like p? Being able to do this allows us to answer questions like

- What is the probability the candidate will be elected?
- What is the probability a toy will contain lead?

- What is the probability that a jar of peanut butter will be contaminated with *Salmonella*?
- What is the chance a vaccine will cause a side-effect? Death?

7 Geometric Random Variable

7.1 Definition

Let us now define another type of random experiment that involves independent trials.

Definition: *Geometric random experiment*

1. Each trial either results in a success (1) or a failure (0).
2. The probability of success p is constant across all trials.
3. The trials are independent.
4. The random variable of interest Y is the number of the trial when the first success happens.

There are two differences with the binomial experiment. The number of trials is not declared before-hand, and the random variable of interest is no longer the number of successes.

The sample space is $S = \{1, 01, 001, 0001, \ldots\}$. It is discrete but infinite (countable) because we cannot say the maximum number of failures we might observe before the first success. The random variable Y is a one-to-one function on S. The range of the random variable Y is

$$Y(S) = \{1, 2, 3, \ldots\}$$

also infinite. In particular, if $p = 0$, then it will take ∞ trials before a success if observed.

It is not hard to see that the

$$p_Y(1) = p \qquad p_Y(2) = (1-p)p \qquad p_Y(3) = (1-p)^2 p$$

etc. Unlike for the Binomial random variable, there is only one way to obtain each value of the random variable (the induced partition is on simple events).

Definition: *geometric random variable*

A discrete random variable Y has a geometric probability distribution if its pmf is

$$p_Y(y) = (1-p)^{y-1}p, \qquad y = 1, 2, \ldots, 0 \leq p \leq 1$$

and we say $Y \sim$ Geometric(p), a distribution with one parameter p.

Can you show that the pmf defined is a valid pmf, i.e. $\sum_y p_Y(y) = 1$?

Is it a pmf?

To verify that the pmf proposed for the Geometric random variable is a pmf, we need to sum over the range $Y(S)$ and obtain 1.

$$\begin{array}{rcll} \sum_{y=1}^{\infty}(1-p)^{y-1}p & = & p\sum_{y=1}^{\infty}(1-p)^{y-1} & \\ & = & p\sum_{x=0}^{\infty}(1-p)^{x} & \text{change of variable } x = y-1 \\ & = & p\frac{1}{1-(1-p)} & \text{Identity: } \sum_{i=0}^{\infty}p^i = \frac{1}{1-p} \text{ if } 0 \leq p < 1 \\ & = & 1 & \end{array}$$

The above works for $p < 1$, but if $p = 1$, then $p_Y(1) = 1$ and $p_Y(y) = 0$ for $y > 1$, so we also verify the result.

Now, I wouldn't propose a pmf that didn't satisfy this requirement, but the exercise demonstrates the utility of the geometric series defined above. I presume you've seen this geometric series before. It's time to dig out and dust off your memories about it.

7.2 Expectation & Variance

Theorem 32. *Given $Y \sim$ Geometric(p),*

$$E[Y] = \frac{1}{p} \qquad\qquad V(Y) = \frac{1-p}{p^2}$$

Proof.

$$\begin{array}{rcll} E[Y] & = & \sum_{y=1}^{\infty} y(1-p)^{y-1}p & \text{definition of expectation and Geometric pmf} \\ & = & -p\sum_{y=1}^{\infty} \frac{d}{dp}(1-p)^y & \text{because } \frac{d}{dp}(1-p)^y = -y(1-p)^{y-1} \\ & = & -p\frac{d}{dp}\left[\sum_{y=1}^{\infty}(1-p)^y\right] & \text{legal to pull the derivative out if } p < 1 \\ & = & -p\frac{d}{dp}\left[\frac{1}{1-(1-p)} - 1\right] & \text{use geometric series result} \\ & = & -p\frac{d}{dp}\left[\frac{1-p}{p}\right] & \text{algebra} \\ & = & -p\left[\frac{-1}{p^2}\right] & \text{derivative} \\ & = & \frac{1}{p} & \end{array}$$

The proof of the formula for $V(Y)$ is obtained the same way, except second derivatives are required. □

7.3 Examples

Example:

Suppose you want to survey people who do not support Obama. You proceed by randomly calling people on the phone. You ask them if they approve of Obama. If they don't, you give them a survey. The current approval rating of the President is 68%. What is the probability that you will find someone to survey within the first three phone calls? What is the expected number of calls until someone is found to be surveyed?

The trial is a call to a person. A *success* is finding someone who does not approve of Obama. We'll assume phone calls are independent, and we'll assume the probability of success $p = 1 - 0.68 = 0.32$ is constant during our experiment. (We'll also not worry about people who don't answer the phone or decline to take the survey.)

If we let Y be the number of phone calls until we find someone to survey, then $Y \sim$ Geometric(0.32). We are asked to find $P(Y \leq 3)$.

$$\begin{array}{rcl} P(Y \leq 3) & = & p_Y(1) + p_Y(2) + p_Y(3) \\ & = & 0.32 + 0.32 \times 0.68 + 0.32 \times 0.68^2 \approx 0.69 \end{array}$$

Notice that

$$\begin{array}{rcl} P(Y \leq y) & = & p + p(1-p) + p(1-p)^2 + \cdots + p(1-p)^{y-1} \\ & = & p\left(1 + (1-p) + (1-p)^2 + \cdots + (1-p)^{y-1}\right) \\ & = & p\dfrac{1-(1-p)^y}{1-(1-p)} \text{see Wikipedia:Geometric series or derive it from the infinite version above} \\ & = & 1-(1-p)^y \end{array}$$

and in this example, $PY \leq 3) = 1 - (0.68)^3 = 0.685568$, as we computed above.

Finding the expected number of calls until you find a person to survey is trivial. From the theorem, we immediately obtain

$$E[Y] = \frac{1}{0.32} = 3.125$$

7.4 Geometric in R

Geometric Distribution in R

There is a complication related to the Geometric distribution in that many people instead define $Z \sim$ Geometric(p), where $Z = Y - 1$ is the *number of failures* before the first success. Because there is a one-to-one mapping from Y to Z or vice versa, it should be clear (they induce the same partition) that

$$p_Y(y) = P(Y = y) = P(Z + 1 = y) = P(Z = y - 1)$$

or, equivalently

$$p_Y(z + 1) = P(Y = z + 1) = P(Z + 1 = z + 1) = P(Z = z)$$

Thus, if you are asked for the probability that there are 3 failures before the first success, you should compute $p_Y(4)$.

Since R uses the random variable Z instead of Y, if you need to compute the probability that the first success occurs at trial 6, then you need to compute the probability that there are 5 failures before the first success using the following functions that return Geometric probabilities.

R function	Probability it computes
`dgeom(x, prob)`	$p_Y(x + 1)$
`pgeom(q, prob)`	$P(Y \leq q + 1)$
`rgeom(n, prob)`	generates realizations of Geometric rvs

where `prob` is the probability of success p.

7.5 Maximum Likelihood Estimation of p

Maximum Likelihood Estimation of p

Suppose we assume $Y \sim$ Geometric(p) but we don't know p. We observe $Y = y_0$. Can we use this information to estimate p? Yes.

Recall the likelihood function is the probability of the observed data

$$L(p) = P(Y = y_0) = p_Y(y_0) = (1 - p)^{y_0 - 1} p$$

It is a function of the unknown parameter p.

With maximum likelihood estimation, we want to choose an estimate of p, call it $\hat{p}$, such that the observed data y_0 has high probability, i.e. is as likely as possible. In other words, we seek $\hat{p}$ that maximizes $L(p)$, or as we argued last time equivalently the $\hat{p}$ that maximizes the log likelihood $l(p)$.

$$l(p) = (y_0 - 1) \ln(1 - p) + \ln p$$

To find the maximum, we take the derivative and set it to 0

$$\begin{aligned} \frac{d}{dp} l(\hat{p}) &= 0 \\ -\frac{y_0 - 1}{1 - \hat{p}} + \frac{1}{\hat{p}} &= 0 \\ -(y_0 - 1)\hat{p} + (1 - \hat{p}) &= 0 \\ \hat{p} &= \frac{1}{y_0} \end{aligned}$$

It is a maximum because $\frac{d^2}{dp^2} l(p) = -\frac{y_0 - 1}{(1-p)^2} - \frac{1}{p} < 0$.

In particular, if $y_0 = 1$, then we estimate $\hat{p} = 1$. If $y_0 = 2$, then we estimate $\hat{p} = \frac{1}{2}$.

8 Negative Binomial

8.1 Definition

Suppose that we are not interested in the first success, rather we desire to learn about the rth success. So, take a Geometrix experiment, but let it proceed until the rth success comes in. Then define Y as the trial when this happens. Obviously, $Y(S) = \{r, r+1, \ldots\}$. Y cannot be less than r because then there could not have been r successes yet.

Suppose $Y = y$. In order for the rth success to occur on the yth trial, we obviously need a success (the rth) to occur on the yth trial. That happens with probability p, and it happens independently of everything else. In addition, in order for the yth trial to give us the rth success, it must be true that $r-1$ successes happened before the yth trial, i.e. in the first $y-1$ trials. What is the probability of getting $r-1$ successes in $y-1$ trials? That is Binomial probability, namely

$$\binom{y-1}{r-1} p^{r-1}(1-p)^{y-r}$$

Putting it all together (multiplying the probabilities of these independent events: success on trial y and $r-1$ success in $y-1$ trials), we have

Definition: *negative binomial*

> A random variable $Y \sim \text{negBinomial}(p, r)$ is said to have a Negative Binomial distribution if it has the following pmf
>
> $$p_Y(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r} \qquad y = r, r+1, r+2, \ldots, 0 \le p \le 1$$
>
> Notice this pmf looks very much like the Binomial pmf, but is a bit different.

8.2 Expectation & Variance

Theorem 33. *Given $Y \sim$ negBinomial(p, r), we have*

$$E[Y] = \frac{r}{p} \qquad V(Y) = \frac{r(1-p)}{p^2}$$

We will not prove these results, but the expectation should be intuitive. If it takes on average $\frac{1}{p}$ to get to the first success (expected value of Geometric random variable), then it will take another $\frac{1}{p}$ to get to the second because the trials are independent, and whatever happened before the first success cannot impact what happens next. Continue this argument and you see the expected wait time until the rth success is $r \times \frac{1}{p}$, the result above.

8.3 Examples

Example:

> Consider oil exploration. If all explorations are independent with constant probability of success $p = 0.2$, then (a) find the probability that the first oil strike occurs on the third exploration, (b) the probability that the third strike occurs on the 7th exploration, and (c) the mean and variance the number of explorations needed to get 3 working wells. Also, if you imagine many companies doing explorations to get 3 wells, and for each you record the number of explorations required to get the third well, then use the empirical rule to find the range within which you expect 95% of these numbers to fall.

(a) $Y \sim$ Geometric(0.2) and we seek $p_Y(3) = 0.8^2 \times 0.2 = 0.128$

(b) $Y \sim$ negBinomial$(3, 0.2)$ and we seek $p_Y(7) = \binom{6}{2} p^3(1-p)^4 \approx 0.049$.

(c) Y is as in (b), and we seek $E[Y] = \frac{3}{0.2} = 15$ and $V(Y) = \frac{3 \times 0.8}{0.2^2} = 60$. By the emprical rule, about 95% of the time we expect the number of explorations to fall in the range $(15 - 2 \times \sqrt{60}, 15 + 2 \times \sqrt{60}) = (-0.49, 30.49)$.

8.4 Negative Binomial in R

As for the Geometric, some authors define the Negative Binomial random variable Z as the *number of failures before* the rth success. Thus, their Z is our $Y - r$. R defines the Negative Binomial as Z and uses $p_Z(z)$ as its pmf. The following relations encode the relationships:

$$p_Z(z) = p_Y(z + r) \qquad\qquad p_Y(y) = p_Z(y - r)$$

Command	Arguments	Probability Computed
`dnbinom(x, size, prob)`	`size`$= r$, `prob`$= p$	$p_Y(x + r)$ or $p_Z(x)$
`pnbinom(q, size, prob)`	`size`$= r$, `prob`$= p$	$P(Y \leq q + r)$ or $P(Z \leq q)$

8.5 More Examples

Example:

Suppose a machine produces parts one at a time and each part is *independently* defective with probability 0.1. Suppose the machine operator needs to produce 100 good parts in order to go home. What is the probability the operator will make a total of 120 parts before he goes home?

We view each part made as a trial. They are independent. With probability 0.9, the outcome is successful: the part is non-defective. Otherwise, the trial is unsuccessful and a defective part is made. Let Y be the number of parts made when the 100th success comes out of the machine. By assumptions $Y \sim$ negBinomial$(0.9, 100)$, so

$$p_Y(100) = \binom{119}{99} 0.9^{100} 0.1^{20} \approx 0.00652$$

as calculated by `dnbinom(x=20, size=100, prob=0.9)`, recalling we need to give R the number of failures, $Z = Y - r = 120 - 100$.

The Negative Binomial random variable has an interesting property that will now be demonstrated. Continuing with our example, suppose that as the operator is about to leave, the boss arrives and tells the operator 10 more good parts are needed. What is the probability that the operator will end up making more than 150 parts throughout the day to meet both orders?

Above, we had $Y \sim$ negBinomial(0.9, 100). Now introduce $X \sim$ negBinomial(0.9,10) the number of total parts the operator produced in order to make 10 additional good parts. We are queried about the sum $X + Y$, the total number of products the operator makes that day.

It turns out that if X and Y are two independent Negative Binomial random variables with the same probability of success p, but different r's, say $r_X = 10$ and $r_Y = 100$ as in our example, then $X + Y \sim$ negBinomial$(0.9, r_X + r_Y)$. Because the trials are independent, X and Y are independent as desired. In this case,

$$P(X + Y > 150) = 1 - P(X + Y \leq 149) \approx 4.70 \times 10^{-9}$$

computed as `1 - pnbinom(q=149-110, size=110, prob=0.9)`.

Lemma 34. *If $X \sim$ negBinomial(p, r_X) and $Y \sim$ negBinomial(p, r_Y) are independent, then $X + Y \sim$ negBinomial$(p, r_X + r_Y)$.*

9 Hypergeometric Distribution

9.1 Definition

Consider an experiment of sampling n independent individuals *without replacement* from a large population and characterizing them as a "success" or "failure." For example, a poll that asks n people a "yes"/"no" question would fall into this category. We are interested in the number of "successes" in our sample.

If the population from which the sample is taken is large, then the Binomial distribution would apply. Let's consider what happens if the population is small, say of finite size N. If there are r individuals in the population who are successes, then before we have sampled anyone, the initial probability of success is

$$p_0 = \frac{r}{N}$$

Now, let's consider what happens upon sampling. There are two choices, depending on whether our first sampled individual is a success, with probability p_0, or a failure, with probability $1 - p_0$. If it is a success, then the 1-sampled population is left with $r - 1$ successes and still $N - r$ failures, so the new probability of success is

$$p_1 = \frac{r-1}{N-1}$$

On the other hand, if the first draw is a failure, then the 1-sampled population is left with r successes and $N - r - 1$ failures, so the new probability of success is

$$p_1 = \frac{r}{N-1}$$

In either case, the probability of success on the next draw has changed, and it will continue to change in a random fashion throughout the sampling process. One of the Binomial experiments assumptions has been violated: the probability of success is not constant across all trials.

Notice, if N is very large, then

$$p_0 \approx p_1$$

This justifies the use of the Binomial distribution when the population from which you sample is very large (more later).

When N is small, we turn to a new kind of experiment.

Definition: *Hypergeometric Experiment*

A random experiment with the following properties is called a *Hypergeometric experiment*.

1. The population is of size N and initially there are r of type "success." The rest are of type "failure."
2. Sample n randomly and without replacement.
3. Record $Y =$ the number of "successes."

Based on your combinatorics experience, you should be able to derive the following pmf without difficulty.

Definition: *Hypergeometric Random Variable*

The random variable $Y \sim \text{Hypergeometric}(N, r, n)$ iff it has pmf

$$p_Y(y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}, \qquad \max(0, n + r - N) \leq y \leq \min(r, y)$$

The imposed limits result because the number of successes in the sample cannot exceed the number of successes in the population: $y \leq r$ and the number of failures in the sample cannot exceed the number of failures in the population: $n - y \leq N - r$.

9.2 Hypergeometric in R

R and the book seem to disagree as much as possible on notation, but it is still easy to use the functions once you learn the argument conventions.

R Command	Arguments	Probability Computed
`dhyper(x, m, n, k)`	m$= r$, n$= N - r$, k$= n$	$p_Y(x)$
`phyper(q, m, n, k)`	m$= r$, n$= N - r$, k$= n$	$P(Y \leq q)$

9.3 Expectation & Variance

Theorem 35. *The random variable $Y \sim$ Hypergeometric(N, r, n) has*

$$E[Y] = \frac{nr}{N} \qquad V(Y) = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$$

We will not discuss a proof. To help you remember these formula, note that $E[Y] = np_0$, where $p_0 = \frac{r}{N}$ is the relative frequency of successes in the whole population. Furthermore, $V(Y) = np_0(1-p_0)\left(\frac{N-r}{N-1}\right)$. Both of these formula look exactly (in the case of expectation) or very close (in the case of variance) to the corresponding results for the Binomial.

Properties:

1. Another important property of the Hypergeometric is

   $$p_{\text{Hypergeometric}}(y) \approx p_{\text{Binomial}}(y)$$

   when N is large. This implies that you can substitute probability calculations for the Hypergeometric with probability calculations from the Binomial. They should match up to increasing numbers of decimal places as N increases. As a test, see the following R code:

```
# large N
> dhyper(10, m=1500, n=2000, k=20)
[1] 0.1436238
> dbinom(10, size=20, prob=1500/(1500+2000))
[1] 0.1433673
# small N
> dhyper(5, m=15, n=20, k=10)
[1] 0.2536151
> dbinom(5, size=10, prob=15/(15+20))
[1] 0.2219864
```

9.4 Examples

Example:

Suppose a radio has 6 transistors, 2 of which are defective. You remove 3 of the transistors. What is the pmf for the number of defective transistors you removed?

Let Y be the number of defective transistors your remove. The range $Y(S) = \{0, 1, 2\}$. $Y(S)$ does not include 3 because $y \leq r$ and $r = 2$. The pmf is

$$p_Y(0) = \frac{\binom{2}{0}\binom{4}{3}}{\binom{6}{3}} \qquad p_Y(1) = \frac{\binom{2}{1}\binom{4}{2}}{\binom{6}{3}} \qquad p_Y(2) = \frac{\binom{2}{2}\binom{4}{0}}{\binom{6}{3}}$$

Example:

Capture-Recapture. The following is a common scheme for estimating the size of a relatively small population. Suppose you capture k animals, tag them all, and then release them back into the population. At some later time, you collect another sample of size n and note the number of tagged animals. To put numbers to this example, consider $k = 4$, $n = 3$, $y = 1$.

The goal is to estimate N. We will use the principle of maximum likelihood. If we assume the number of tagged animals in our second capture is $Y \sim$ Hypergometric$(N, 4, 3)$, then the probability of our observation is

$$\begin{aligned} p_Y(1) &= \frac{\binom{4}{1}\binom{N-4}{2}}{\binom{N}{3}} \\ &= \frac{4(N-4)(N-3)3!}{2!N(N-1)(N-2)} \end{aligned}$$

This is the likelihood $L(N)$ and we could go about solving for the N that maximizes it, but we can also try a few values by brute force (see `R` commands below). We conclude that good estimates of N are $\hat{N} = 11$ or $\hat{N} = 12$, as both give the same maximal probability of 0.51 of observing the data $Y = 1$.

```
> dhyper(1, m=4, n=9-4, k=3)
[1] 0.4761905
> dhyper(1, m=4, n=10-4, k=3)
[1] 0.5
> dhyper(1, m=4, n=11-4, k=3)
[1] 0.5090909
> dhyper(1, m=4, n=12-4, k=3)
[1] 0.5090909
> dhyper(1, m=4, n=13-4, k=3)
[1] 0.5034965
```

10 Poisson Distribution

10.1 Introduction

One of the most useful discrete random variables is the Poisson random variable. We will start by defining it, and then provide motivation.

Definition: *Poisson random variable*

Random variable $Y \sim$ Poisson(λ) is said to have a Poisson distribution iff its pmf is

$$p_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!}, \qquad y = 0, 1, 2, \ldots, \lambda > 0$$

Notice the range of a Poisson random variable are the counting numbers $Y(S) = \{0, 1, 2, \ldots\}$, which reflects this distributions value as a distribution for the count of random events occurring in time or space.

Figure 2: Example of Poisson Process

First, we'll verify that the proposed pmf is a proper probability distribution.

$$\begin{aligned} P[y \in Y(S)] &= \sum_{y \in Y(S)} p_Y(y) \\ &= \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} \\ &= e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\ &= e^{-\lambda} e^{\lambda} = 1 \end{aligned}$$

You should recognize the use of the Taylor series expansion for e^{λ}.

The Poisson distribution arises out of the so-called "Law of Rare Events." Consider events occurring sporadically in space or time or space$\times$time. Figure 2 shows a total of 4 events occurring between time 0 and time T. The events can be anything, such as: (1) people arriving at a checkout in a department store, (2) airplanes taking off from Des Moines International airport, (3) trains passing through Ames, (4) alpha particles being emitted from radioactive nuclei, (5) earthquakes striking along a fault line in time, (6) mutations occurring along a genome, (7) phone calls arriving on your cell phone, (8) meteors striking the earth's surface, (9) snowflakes forming in the atmosphere, etc. The point is that the events are relatively rare, such that no two events will occur at the exact same moment in time or space. The random experiments of the type just described are called *Poisson Processes*.

It turns out that under very general conditions, if the events occur independently of each other, then the total number of events in any interval of time or space Y is distributed as a Poisson random variable with λ equal to the expected number of events. We will now prove this assertion, using Fig. 2 as our example.

Proof. Divide $[0, T]$ into n small equi-sized subintervals, choosing n large enough so that the number of events in any subinterval should almost never exceed 1. If Y_i is the number events occurring in the ith subinterval, then

$$\begin{aligned} P(Y_i > 1) &\approx 0 \\ P(Y_i = 1) &\approx p \\ P(Y_i = 0) &\approx 1 - p \end{aligned}$$

We know that $P(Y_i = 1)$ is the same constant p for all subintervals i because each subinterval is of the same size and events are occurring independently of each other.

Under our assumptions, the random variable Y is the total number of subintervals that contain events, but now we recognize Y as a Binomial random variable with number of trials n and probability of success p. Thus,

$$\begin{aligned} P(Y = y) &= \tbinom{n}{y} p^y (1-p)^{n-y} && \text{definition of Binomial rv} \\ &= \tfrac{n(n-1)(n-2)\cdots(n-y+1)}{y!} \left(\tfrac{p}{1-p}\right)^y (1-p)^n && \text{rearrangement} \end{aligned}$$

Now, since event locations are random and we must divide the interval $[0, T]$ up *before* we observe the events, we cannot actually guarantee $P(Y_i > 1)$ unless $n \to \infty$. By dividing $[0, T]$ into subintervals we

are not changing the random experiment, so the expectation of Binominal Y, namely $E[Y] = np$, is not changing. Let $\lambda = np$, so it is clear that as $n \to \infty$, $p \to 0$. These imply

$$\begin{aligned} n(n-1)(n-2)\cdots(n-y+1) &\to n^y \\ (1-p)^n &\to 1 \\ (1-p)^n = \left(1 - \frac{\lambda}{n}\right)^n &\to e^{-\lambda} \end{aligned} \tag{2}$$

where the last limit is probably something you saw in the calculus of limits. If you don't remember, try writing out the binomial expansion of $\left(1 - \frac{\lambda}{n}\right)^n$ and looking for the Taylor's series expansion of $\exp()$ by using eq 2. Putting all these limits into our equation yields

$$P(Y = y) = \frac{n^y}{y!} p^y e^{-\lambda} = \frac{\lambda^y}{y!} e^{-\lambda}$$

□

An implication of the above proof is

Properties:

1. $p_{\text{Binomial}}(y) \approx p_{\text{Poisson}}(y)$ with $\lambda = np$ as n gets large and p gets small. A rule of thumb is n large, p small and $np \leq 7$.
2. If $Y_1 \sim \text{Poisson}(\lambda_1)$ is independent of $Y_2 \sim \text{Poisson}(\lambda_2)$, then $Y_1 + Y_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.
3. If $Y \sim \text{Poisson}(\lambda)$ and events are independently labeled as successes with probability p, then X the number of *successful* events is also Poisson with mean $\lambda_X = p\lambda$.

10.2 Poisson with R

R Function	Arguments	Probability Computed
`dpois(x, lambda)`	`lambda`=λ	$p_Y(x)$
`ppois(q, lambda)`	`lambda`=λ	$P(Y \leq q)$

10.3 Expectation & Variance

One of the defining characteristics of the Poisson random variable is that the expectation equals the variance. In fact, people often test this property when checking to see if their data follow the Poisson distribution (for example to test for independence of events).

Theorem 36. *If $Y \sim$ Poisson(λ), then*

$$E[Y] = V(Y) = \lambda$$

Proof. The proof for the expectation is a simpler example of the proof for the variance, so we'll skip it. Also, I had claimed that the proof of the Binomial variance was obtained in this way, so see this as a demonstration of the technique I motivated, but did not demonstrate in the Binomial unit.

$$\begin{aligned} E[Y(Y-1)] &= \sum_{y=0}^{\infty} y(y-1)\frac{e^{-\lambda}\lambda^y}{y!} && \text{definition of expectation} \\ &= \sum_{y=2}^{\infty} y(y-1)\frac{e^{-\lambda}\lambda^y}{y!} && \text{first two summands are } 0 \\ &= \lambda^2 \sum_{y=2}^{\infty} \frac{e^{-\lambda}\lambda^{y-2}}{(y-2)!} && \text{cancel } y(y-1) \text{ and pull out } \lambda^2 \\ &= \lambda^2 \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{x!} && \text{change of variable } x = y-2 \\ &= \lambda^2 && \text{pmf sums to 1 over } X(S) \end{aligned}$$

Then $E[Y(Y-1)] = E[Y^2] - E[Y] = E[Y^2] - \lambda^2 + \lambda^2 - \lambda = V(Y) + \lambda^2 - \lambda$ so

$$V(Y) = \lambda$$

□

10.4 Examples

Example:

1. If a fire alarm indicates a fire with probability 0.03, what is the probability the the fire department will have to put out at least one fire after responding to 100 fire alarms?

   Let Y be the number of fires, then Y is the number of successes in $n = 100$ trials with success probability $p = 0.03$, but we notice that n is large, p is small and $np = 3 \leq 7$, so we can use the Poisson approximation. The mean of this Poisson distribution is $\lambda = np = 3$, so

   $$P(Y \geq 1) = 1 - P(Y = 0) = 1 - e^{-3} \approx 0.95$$

2. Suppose you attend class and forget to turn off your cell phone. If your receive about 5 calls during the 12-hr waking hours of the day, and the class lasts 1 hour, what is the chance you will get a call during the class?

   We should think Poisson process because we are given a rate, 5 calls per 12 hours, and rates should make us think Poisson. Also, we are interested in the number of calls occurring in a 1-hour time period, so the interval of interest is 1 hour. We need λ, the expected number of calls during 1 hour.

   $$\lambda = \frac{5 \text{ calls}}{12 \text{ hours}} = \frac{5}{12} \text{ calls per hour}$$

   The probability we seek is

   $$1 - P(Y = 0) = 1 - e^{-5/12} = 0.3407594$$

Framework for Solving a Poisson Problem

1. Identify if the problem is asking about a Poisson random variable.
   (a) If it looks like a Binomial question, check if n is large and p is small. According to Wikipedia, there are two conditions when a Binomial experiment can be reduced to a Poisson process:
      i. $n \geq 20$ and $p \leq 0.05$, or
      ii. $n \geq 100$ and $np \leq 10$
   (b) If a rate is provided (e.g. 3 fire alarms per day), think Poisson.
2. Identify interval in space, time or space$\times$time addressed in the question. Find the size of this interval. For 1D, you need a length, for 2D you need an area, for 3D you need a volume, for 4D (e.g. time$\times$space), you need a 4D volume.
3. Compute mean number of events λ in interval.
   (a) If you converted from Binomial problem, then $\lambda = np$.
   (b) If given a rate, you may need to adjust for your interval size. For example, if the interval is 1 hour, then $\lambda = 3/24$ fire alarms per hour.
4. Then the number of events in the interval $Y \sim$ Poisson(λ). Use pmf, $E[Y]$, and $V(Y)$ to answer question.

It could be a good idea to derive such a framework for answering any word problem from this chapter.

10.5 Poisson Properties

There are two properties of Poisson random variables that can be particularly helpful, and also partly explain the ubiquity of this distribution.

Properties: Poisson process

1. If $Y_1 \sim \text{Poisson}(\lambda_1)$ and $Y_2 \sim \text{Poisson}(\lambda_2)$ are independent, then
$$Y_1 + Y_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$$
2. If $Y \sim \text{Poisson}(\lambda)$ and events are marked as "successful" independently with probability p, then if X is the number of successful events, then
$$X \sim \text{Poisson}(p\lambda)$$

10.6 Advanced Examples

Example:

1. Suppose a parking lot has two entrances I and II and cars arrive at entrance I at a rate of 3 per hour and independently at entrance II at a rate of 4 per hour. What is the probability that 3 cars arrive in 1 hour?
   If $Y_1 \sim \text{Poisson}(3)$ is the number of cars that arrive through entrance I in an hour and $Y_2 \sim \text{Poisson}(4)$ is the number of cars that arrive through entrance II in an hour, then we seek $P(Y_1 + Y_2 = 3)$. Using property 1, we know $Y_1 + Y_2 \sim \text{Poisson}(7)$, so
$$P(Y_1 + Y_2 = 3) = \frac{e^{-7}7^3}{3!} \approx 0.05$$
   Now, we use this example to prove property 1.

   Proof. Notice that the pair $(Y_1, Y_2) \in \{(0,3),(1,2),(2,1),(3,0)\}$ when $Y_1 + Y_2 = 3$, so this set partitions the event $Y_1 + Y_2 = 3$. Thus, we can use the addition law for mutually exclusive events to find

$$\begin{array}{rcll} P(Y_1+Y_2=3) & = & \sum_{y_1,y_2:y_1+y_2=3} P(Y_1=y_1, Y_2=y_2) & \text{addition rule} \\ & = & \sum_{y_1,y_2:y_1+y_2=3} P(Y_1=y_1)P(Y_2=y_2) & \text{independence of } Y_1 \text{ and } Y_2 \\ & = & \sum_{y_1,y_2:y_1+y_2=3} \frac{e^{-3}3^{y_1}}{y_1!}\frac{e^{-4}4^{y_2}}{y_2!} & Y_1 \text{ and } Y_2 \text{ are Poisson} \\ & = & e^{-(3+4)}\sum_{y_1,y_2:y_1+y_2=3} \frac{3^{y_1}4^{y_2}}{y_1!y_2!} & \text{pull out two constants} \\ & = & \frac{e^{-(3+4)}}{3!}\sum_{y_1,y_2:y_1+y_2=3} \frac{3!}{y_1!y_2!}3^{y_1}4^{y_2} & \text{multiply/divide by 3!} \\ & = & \frac{e^{-(3+4)}}{3!}\sum_{y_1=0}^{3}\binom{3}{y_1}3^{y_1}4^{3-y_1} & \text{rearrangement and } y_2 = 3-y_1 \\ & = & \frac{e^{-(3+4)}}{3!}(3+4)^3 & \text{rearrangement and } y_2 = 3-y_1 \\ & = & \frac{e^{-7}}{3!}7^3 & \text{voila!} \end{array}$$

   Isn't it amazing how that works out so neatly? □
2. Suppose an instructor makes errors on homeworks like a Poisson process with mean 4 errors per homework. Suppose that with constant probability $p = 0.1$ each error may independently be a mathematical error. What is the probability of at least 1 mathematical error on a homework?
   If Y is the number of errors and X is the number of mathematical errors, then property 2 gives us that
$$X \sim \text{Poisson}(0.1 \times 4 = 0.4)$$

Thus, the probability of at least one error is

$$1 - P(X = 0) = 1 - e^{-0.4} \approx \frac{1}{3}.$$

Let's use this example to prove property 2.

Proof. If we knew the total number of errors on the homework Y, then $X \mid Y$ would have a Binomial distribution ($n = Y$ independent trials, constant probability of success $p = 0.1$). In other words, the conditional probability

$$P(X = x \mid Y = y) = \binom{y}{x} 0.1^x 0.9^{y-x}$$

Thus, if we only knew Y, then we would know the distribution of X. A neuron should be itching in your brain. We used this phraseology before when discussing the Law of Total Probability.

$$\begin{array}{rcll}
P(X = x) & = & \sum_{y=x}^{\infty} P(X = x \mid Y = y) P(Y = y) & \text{LTP; notice } y \geq x \\
& = & \sum_{y=x}^{\infty} \binom{y}{x} 0.1^x 0.9^{y-x} \frac{e^{-4} 4^y}{y!} & Y \sim \text{Poisson}(4) \text{ and } X \sim \text{Binomial}(y, 0.1) \\
& = & \sum_{y=x}^{\infty} \frac{1}{x!(y-x)!} 0.1^x 0.9^{y-x} e^{-4} 4^y & \text{cancel } y! \\
& = & \sum_{y=x}^{\infty} \frac{1}{x!(y-x)!} (0.1 \times 4)^x (4 \times 0.9)^{y-x} e^{-4} & \text{distribute } 4^x \\
& = & \frac{0.4^x}{x!} \sum_{y=x}^{\infty} \frac{1}{(y-x)!} [4(1 - 0.1)]^{y-x} e^{-4[0.1+(1-0.1)]} & \text{move out constants, } 0.9 = 1 - 0.1 \\
& = & \frac{0.4^x e^{-0.4}}{x!} \sum_{y=x}^{\infty} \frac{1}{(y-x)!} [4(1 - 0.1)]^{y-x} e^{-4(1-0.1)} & \text{move out constant } e^{-4 \times 0.1} \\
& = & \frac{0.4^x e^{-0.4}}{x!} \sum_{z=0}^{\infty} \frac{1}{z!} [4(1 - 0.1)]^{z} e^{-4(1-0.1)} & \text{change of variable } z = y - x \\
& = & \frac{0.4^x e^{-0.4}}{x!} & \text{recognize } Z \sim \text{Poisson}(4[1 - 0.1]) \\
& \sim & \text{Poisson}(0.4) &
\end{array}$$

Another neat result! □

3. Finally, I want to remind you that we started this chapter with functions of random variables $X = f(Y)$, so I can define functions of any random variable that we've defined. Take, for example, the following question.

   Suppose customers arrive independently at a checkout with rate 7 per hour. If it takes 10 minutes to serve each customer and there are enough employees and service stations so that there are no waits, what is the expectation and variance of the total service time for all customers arriving in an hour?

   We think Poisson random variable because we are given a rate of 7 customers per hour. The interval of time that is our interest is one hour and the mean is $\lambda = 7$. Let Y be the number of customers arriving in one interval, then the service time is $X = 10Y$. So,

$$E[X] = E[10Y] = 10E[Y] = 70 \quad V(X) = V(10Y) = 100V(Y) = 700.$$

11 Review

Pop quiz. Identify the discrete random variables discussed in each problem. (Working through these questions would be additional good practice for the exam, but here I will just identify the random variable of interest.)

1. A preacher collects 100 one-dollar bills in a hat he passes around the church. Suppose you contribute 3 bills. What is the probability that the when the hat returns to the preacher and he reaches in and takes out two dollars, one in each hand, that he is holding two of *your* dollar bills in his hands?

   Hypergeometric

Binomial might be a good approximation since $n = 100$ is fairly large.

2. I try starting my car once every minute until it starts. Presuming each attempt is independent and the probability the car starts on each attempt is constant, how many minutes will it take me to start my car?

   Geometric

3. Your genome consists of 3.4 billion nucleotides. The probability that any single nucleotide mutates is 0.000000025. What is the probability that you are a mutant?

   Poisson

   Binomial is exact, Poisson is approximate. The Binomial would not be computable on some calculators, though `R` handles it fine.

4. Two eggs in an egg carton are cracked. I sample eggs without replacement until I find the first one cracked. When can I expect to find a cracked egg?

   No named distribution

   It would be geometric if I sampled with replacement.

5. I play rounds of black jack until I win my second hand. How much can I expect to win or lose if it costs \$5 dollars to play each game and the house has an 8% advantage?

   Negative Binomial

   I'm waiting until the second success. By the way, an 8% advantage for the house means that you lose on average 8% of every bet you make, so you're losing an average $5 \times 0.08 = 0.4$ per hand.

Part III

Continuous Random Variables

12 Introduction

Review: Random Variable

Recall the definition of a random variable.

Definition: *random variable*

> Given a random experiment with sample space S consisting of simple events ω, a random variable X is a function that maps $\omega \in S$ to the real numbers $\Re$.
>
> $$X : \omega \to \Re$$

The *range* $X(S)$ of the random variable is the collection of numbers in $\Re$ that the random variable maps to.

A *discrete random variable* is one where $X(S)$ is countable, i.e. there is a one-to-one map between the elements of $X(S)$ and the positive integers. In other words, one can imagine labeling each simple event $\omega \in X(S)$ with a *unique* integer.

Now we consider cases where $X(S)$ is uncountable, i.e. there are too many numbers to label each one with a unique integer. Any line segment, e.g. $[a, b)$, $[a, b]$, $(a, b]$, $[a, \infty)$, (a, ∞), $(-\infty, b]$, or $(-\infty, b)$, for finite $a, b \in \Re$, is uncountable (Real analysis covers these ideas in detail.). We are talking about *continuous random variables*, although we will be more precise with our definition later.

Examples of such random variables are

- The time until the bus arrives at the bus stop. $X(S) = [0, \infty)$.
- The probability a person will die from heart disease, each probability depends on a person's genes, life experience, exposures, etc. $X(S) = [0, 1]$.
- The distance a discus lands from the discus thrower. $X(S) = [0, \infty)$.
- The distance a dart lands from the bulls-eye given that it does not hit the bulls-eye but hits the dart board. $X(s) = (r, R]$, where r is the radius of the bulls-eye and R is the radius of the board.

12.1 Probability Mass Function Does Not Exist

Previously, we argued that the probability mass function (pmf) $p_X(x) = P(X = x)$ completely defines a discrete random variable. Everything to know about a discrete random variable is summarized in the pmf.

So, we would like to define a pmf for these new random variables that exist on intervals. Unfortunately, it is not possible. We will show this by contradiction.

Suppose we could define $P(X = x)$ for all $x \in X(S)$. Then, it must be true (by the Axioms of Probability) that

$$\sum_{x \in X(S)} P(X = x) = 1 \tag{3}$$

Consider the subset of $X(S)$ whose probability exceeds $\frac{1}{n}$:

$$A_n = \left\{ x : P(X = x) > \frac{1}{n} \right\}$$

Then, the sum over just this subset is

$$\sum_{x \in A_n} P(X = x) > \sum_{x \in A_n} \frac{1}{n} = \frac{|A_n|}{n}$$

where $|A_n|$ is the number of elements in A_n. The above "sub-sum" only converges if A_n is finite: $|A_n| < \infty$. Thus, for all n, $|A_n| < \infty$ in order for eq 3 to converge. Furthermore, the range $X(S)$ is a countable union of A_n:

$$X(S) = A_2 \cup A_3 \cup \cdots = \cup_{n=2}^{\infty} A_n$$

A countable union of finite sets is countable. Thus, we have reached a contradiction. Either $X(S)$ is countable, in which case X is a discrete random variable, or $P(X = x) = 0$ for all but a countable subset of $X(S)$, which again makes X finite.

The main conclusions are the pmf does not exist and $P(X = x) = 0$.

12.2 Probability Distribution

12.2.1 Cumulative Density Function (cdf)

We'll have to come up a novel way to describe random variables of this type.

Definition: *cumulative distribution function (cdf)*

For any random variable Y, the cumulative distribution exists is defined as

$$F(y) = P(Y \leq y), \quad -\infty < y < \infty$$

Notice, the cdf is defined for all possible values of $y \in \Re$.

Properties: cdf

1. The cdf is defined for discrete and continuous random variables.
2. $\lim_{y\to-\infty} F(y) = \lim_{y\to-\infty} P(Y \leq y) = 1 - P(Y \in Y(S)) = 0$.
3. $\lim_{y\to-\infty} F(y) = \lim_{y\to-\infty} P(Y \leq y) = P(Y \in Y(S)) = 1$.
4. For $y_1 < y_2$, $F(y_1) \leq F(y_2)$, so $F(y)$ is a monotonic, non-decreasing function of y.
5. For $y_1 < y_2$, $P(y_1 < Y \leq y_2) = P(Y \leq y_2) - P(Y \leq y_1) = F(y_2) - F(y_1)$.
6. $F(y)$ is right continuous, i.e. $\lim_{y\to y_0^+} F(y) = F(y_0)$ for all $y_0 \in (-\infty, \infty)$.

Definition: *continuous random variable*

A continuous random variable is one for which $F(y)$ is continuous (and with derivative defined at all but a finite number of points).

12.2.2 Probability Density Function (pdf)

Definition: *probability density function (pdf)*

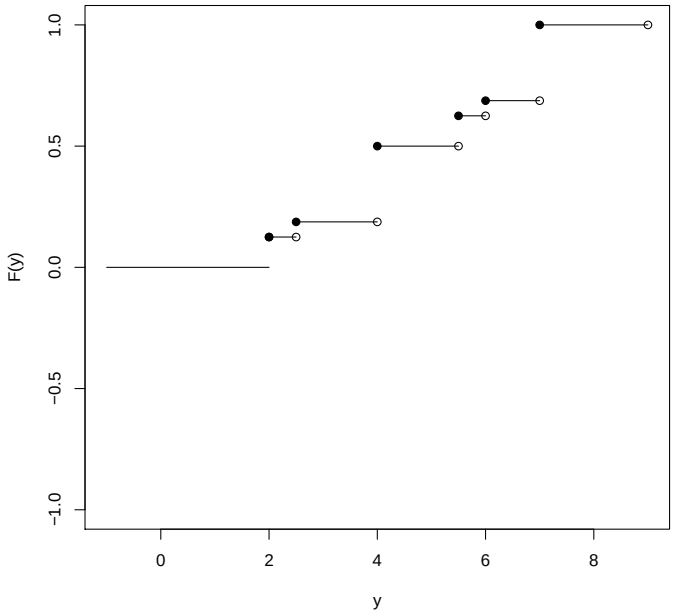
For a continuous random variable, the probability density function if it exists is

$$f(y) = \frac{dF(y)}{dy}$$

Properties: pdf

1. When the pdf exists, $F(y) = \int_{-\infty}^{y} f(t)dt$.
2. $f(y) \geq 0$ because $F(y)$ is non-decreasing.
3. The pdf is *not* a probability, i.e. it is possible for $f(y) > 1$.
4. $\int_{-\infty}^{\infty} f(t)dt = 1$ because $F(y) \to 1$ as $y \to \infty$.

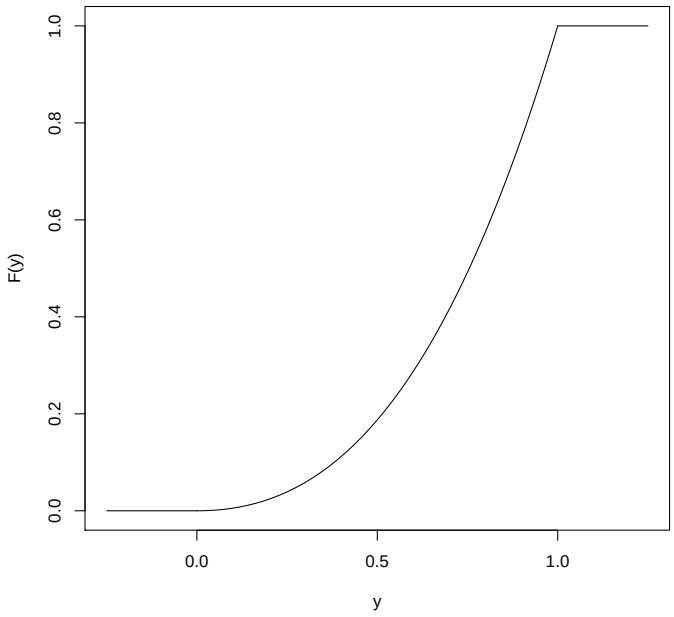
12.2.3 Examples



An example of the cdf of a discrete random variable (in graphical form and as formula).

$$F(y) = P(Y \leq y) = \begin{cases} 0, & \text{for } y < 2 \\ 1/8, & \text{for } 2 \leq y < 2.5 \\ 3/16, & \text{for } 2.5 \leq y < 4 \\ 1/2, & \text{for } 4 \leq y < 5.5 \\ 5/8, & \text{for } 5.5 \leq y < 6 \\ 11/16, & \text{for } 6 \leq y < 7 \\ 1, & \text{for } y \geq 7 \end{cases}$$

An example of a continuous random variable cdf (graphical and formula).



$$F(y) = \begin{cases} \frac{y^3}{2} + \frac{y^2}{2}, & 0 \le y \le 1 \\ 0, & \text{otherwise} \end{cases}$$

What is the pdf?

$$f(y) = \frac{3y^2}{2} + y, \quad 0 \le y \le 1$$

Example: continuous random variable

Suppose you know that waist-to-hip ratio affects life expectancy. In particular, if Y is life expectancy and w is waist-to-hip ratio, then the pdf of life expectacy is given by (made up, but roughly matching life tables)

$$f(y) = C\left[0.0075e^{-100y} + e^{\frac{0.07(y-120)}{w}}\right], \quad y \ge 0$$

where C is chosen such that $f(y)$ is a pdf. Plots of $f(y)$ are given in Fig. 3.

Questions that you should be able to answer:

- If $w = 0.7$, what is the value of C?
- What is the cdf $F(y)$?
- What is $P(Y > 65 + c)$ for some constant c, i.e. that someone lives c years past retirement.
- What is $P(50 \le Y \le 70)$?
- What is $P(Y > 80 \mid Y > 65)$?

12.2.4 Relation of cdf & pdf

Our immediate goal is to gain a better understanding of the meaning of the probability density function (pdf) $f(x)$, for example the one shown in Fig. 4 We have already established that $f(x) \neq P(X = x)$. In fact, $f(x) > 1$ is possible and $P(X = x) = 0$.

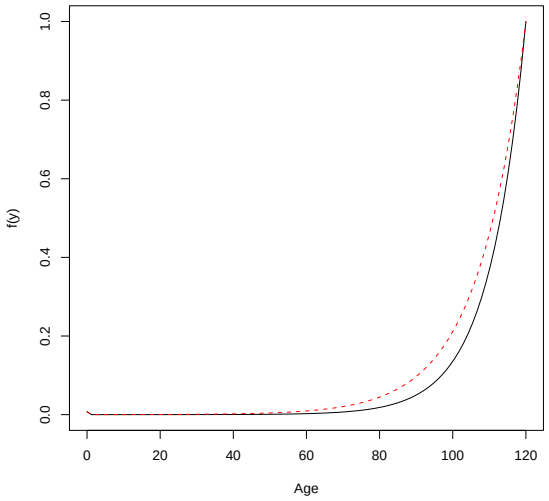


Figure 3: Probability distribution function (pdf) of female life expectancy. Solid line is for waist-to-hip ratio of 0.9 and dotted line is for waist-to-hip ratio of 0.7 (health-wise ideal).

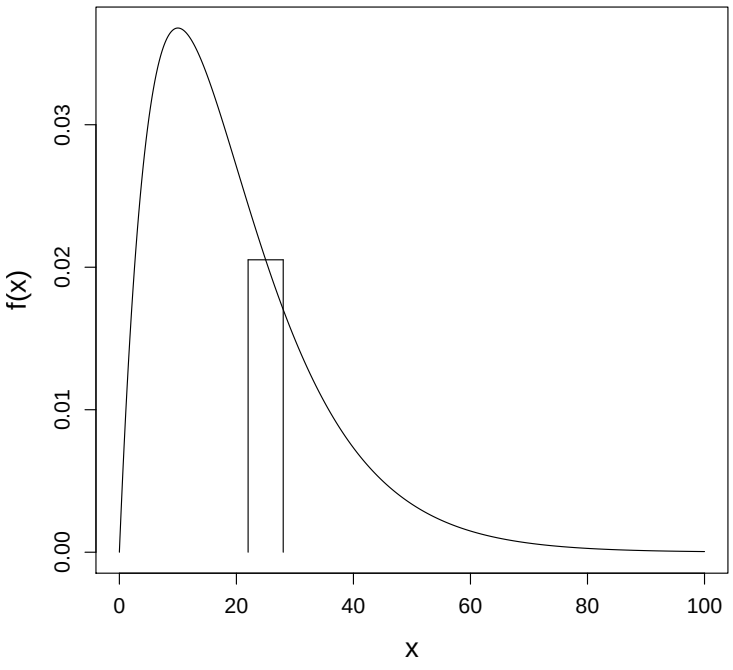


Figure 4: A probability density function.

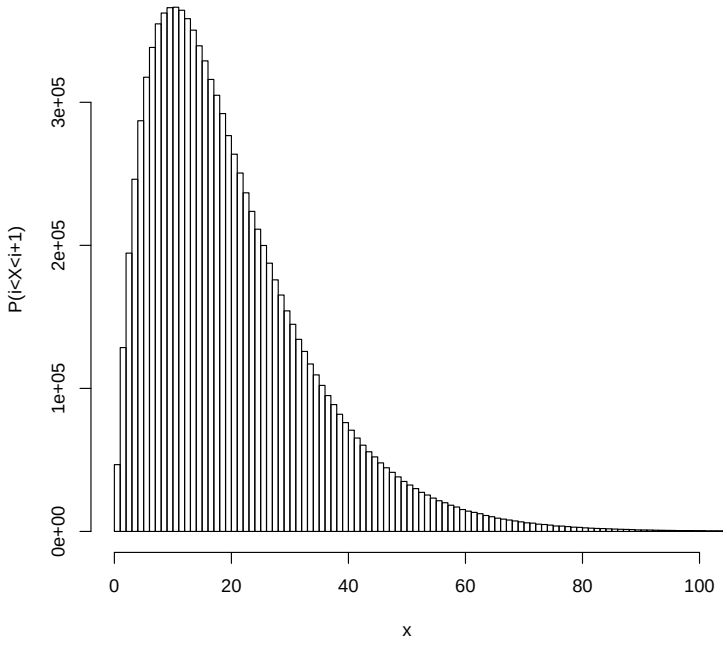


Figure 5: Discrete approximation to Gamma distribution.

What we do know is that $f(x)$, when it exists, is the derivative of the cdf:

$$\begin{array}{rcll} f(x) & = & \frac{dF(x)}{dx} & \text{definition of pdf} \\ f(x) & = & \lim_{dx \to 0} \frac{F(x+dx/2)-F(x-dx/2)}{dx} & \text{one definition of derivative} \\ f(x)dx & \approx & F(x+dx/2)-F(x-dx/2) & \text{for small enough } dx \end{array}$$

On the left, we have $f(x)dx$, the area of a rectangle centered on x, of width dx and of height $f(x)$, for example the rectangle shown in Fig. 4. On the right, we have the probability that random variable X will fall within the small interval $(x - dx/2, x + dx/2]$, as see by utilizing the definition of cdf:

$$\begin{array}{rcl} F(x+dx/2)-F(x-dx/2) & = & P(X \leq x+dx/2)-P(X \leq x-dx/2) \\ & = & P(x-dx/2 < X \leq x+dx/2). \end{array}$$

Thus, $f(x)$ is proportional to the probability that X falls in a tiny interval surrounding x

$$f(x) \propto P(x - dx/2 < X \leq x + dx/2)$$

and we no longer think about exact equality $X = x$, rather tiny intervals $X \in (x - dx/2, x + dx/2]$.

A slightly different way to envision this relationship, is to imagine a discrete approximation to the $f(x)$ curve (see Fig. 5). This approximation is obtained by breaking the range of X, in this case $X(S) = (0, \infty)$ into many equal-width intervals, with boundaries $x_0, x_1, x_2, x_3, \ldots$ such that $x_0 = 0$ and $x_{i+1} - x_i = dx$. To make this work the width dx should be small. In Fig. 5, the width is 1 (not so small, but visible to your eyes). This discretization might remind you of Reimann sums in calculus (yes, that stuff has use outside of calc!). Each rectangle has a height $f(x_0), f(x_1), \ldots$, and the rectange area $f(x_i)dx$ approximates the probability X falls in $(x_i - dx/2, x_i + dx/2]$, i.e. $P(x_i - dx/2 < X \leq x_i + dx/2)$.

A Note About Intervals

How does the probability X falls in $(a, b]$ differ from the probability X falls in (a, b), i.e. with and without inclusion of the endpoint b?

A little thought should make it clear that these probabilities are the same, and both equal to

$$F(b) - F(a)$$

The reason is that $(a, b] = (a, b) \cup b$, where the two events on the right are mutually exclusive, thus

$$P(X \in (a, b]) = P(X \in (a, b)) + P(X = b) = P(X \in (a, b))$$

becuase $P(X = b) = 0$. With continuous random variables *it is OK to get sloppy* and use $<$ and $\leq$ or $>$ and $\geq$ interchangeably.

12.3 Definitions

12.3.1 Quantile/Percentile

Definition: *quantile*

> Given a r.v. Y and $0 < p < 1$, then the p^{th} quantile of Y, denoted ϕ_p, is the smallest value such that
>
> $$P(Y \leq \phi_p) = F(\phi_p) \geq p$$
>
> If Y is continuous, then ϕ_p is the smallest value such that
>
> $$F(\phi_p) = p$$

Why do we know there is ϕ_p such that $F(\phi_p) = p$ when Y is continuous? Because $F(y)$ is continuous for a continuous r.v. by definition.

The *percentile* is the *quantile* multiplied by 100 to discuss it as a percent.

[Insert graphical demonstration.]

Finding quantiles in R

The `q*` functions, like `qbinom`, `qgeom`, `qnbinom`, `qhyper`, and `qpois` are used to find quantiles. The first argument is the probability p you see, e.g. $0.1, 0.25, 0.5, 0.90, 0.95, 0.975$. The remaining arguments are the parameters of the distribution.

Examples.

- Find the 0.2 quantile for $Y \sim$ Poisson(3) distribution. `qpois(0.2, lambda=3)` returns `2`.
- Find the $\phi_{0.05}$ and $\phi_{0.95}$ in order to construct the 90% confidence interval $(\phi_{0.05}, \phi_{0.95})$ in which you expect $Y \sim$ negBinomial$(0.1, 3)$ to fall with at least 90% probability. `qnbinom(0.05, size=3, prob=0.2)` returns `2` and `qnbinom(0.95, size=3, prob=0.2)` returns `27`, yielding range $(5, 30)$, after we add the 3 successes to get the trial numbers (remember `R` returns the number of failures before the rth success).
- Find the largest y_0 such that at least 30% of $Y \sim$ Hypergeometric$(200, 30, 10)$ are expected to be greater than this number. `qhyper(0.7, m=170, n=30, k=10, lower.tail=F)` returns $\phi_{0.7} = 9$, but this guarantees only $P(Y \leq y_0) \geq 0.7$, in fact `phyper(9, m=170, 30, k=10)` $= 0.84$ is strictly greater than 0.7, so $y_0 = \phi_{0.7} - 1 = 8$.

12.3.2 Expectation/Variance

Definition: *expectation*

Given a continuous random variable Y, the expectation is

$$E[Y] = \int_{-\infty}^{\infty} yf(y)dy$$

when it exists, meaning $\int_{-\infty}^{\infty} |y|f(y)dy < \infty$.

Theorem 37.

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy$$

$$E[c] = \int_{-\infty}^{\infty} cf(y)dy = c\int_{-\infty}^{\infty} f(y)dy = c$$

$$E[cg(Y)] = \int_{-\infty}^{\infty} cg(y)f(y)dy = c\int_{-\infty}^{\infty} g(y)f(y)dy = cE[g(Y)]$$

$$E\left[\sum_{i=1}^{n} a_i g_i(Y)\right] = \int_{-\infty}^{\infty} \sum_{i=1}^{n} a_i g_i(y)f(y)dy = \sum_{i=1}^{n} \int_{-\infty}^{\infty} a_i g_i(y)f(y)dy = \sum_{i=1}^{n} a_i E[g_i(Y)]$$

Definition: *variance*

Given a continuous r.v. Y, the variance is

$$V(Y) = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy$$

Properties:

1. $V(Y) = E[Y^2] - (E[Y])^2$, with basically same proof as discrete r.v. (see homework).
2. $V(aY+b) = a^2V(Y)$, again proved as for discrete.

13 Uniform Distribution

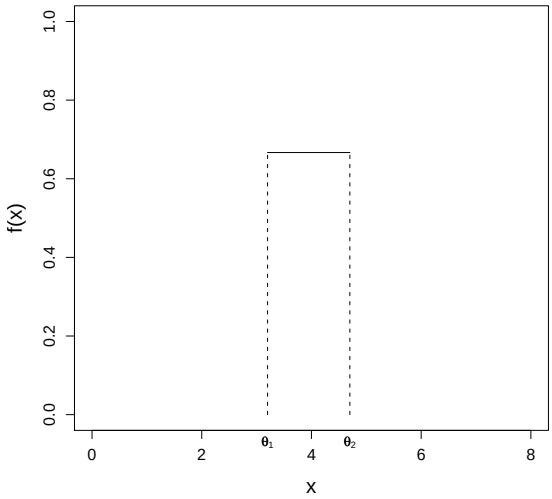
13.1 Probability Density Function

Definition: *uniform distribution*

For $\theta_1 < \theta_2$, the continuous random variable $Y \sim \text{Uniform}(\theta_1, \theta_2)$ is said to have a *uniform distribution* on interval (θ_1, θ_2) if and only if its pdf is

$$f(y) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & \theta_1 \le y \le \theta_2 \\ 0, & \text{otherwise} \end{cases}$$

For example, the following is a plot of the pdf for Uniform$(3.2, 4.7)$.



Notes.

- The uniform distribution has 2 parameters, θ_1 and θ_2 that define the lower and upper limit of the range of Y.
- All allowable values of the random variable, i.e. all points in (θ_1, θ_2), are equally probable.
- The uniform distribution is important for two reasons:
  - **Random number generation.** If a computer program can generate $U \sim \text{Uniform}(0, 1)$ (e.g. `runif(1)`), then $Y = F^{-1}(U) \sim \text{F}(y)$. In other words, once a computer can generate a uniform random variables, it can also generate random variables Y from any distribution $F(y)$ if the inverse function $F^{-1}(\cdot)$ is available.
  - **Many physical phenomena have approximate uniform distribution.** If events occur as a Poisson process, then *given an event has occurred* in the interval (a, b), the exact time or location of that event has a Uniform(a, b) distribution.
- $F(y) = \int_a^y \frac{1}{b-a} dt = \frac{t}{b-a}\Big|_a^y = \frac{y-a}{b-a}$.
- The *standard uniform* random variable is $Y \sim \text{Uniform}(0, 1)$, with $f(y) = 1$.

Relation to Poisson

To be more concrete about the relationship to the Poisson, we will prove the following theorem about the time/location of an event after it is known that an event has occurred. (Recall, that the Poisson random variable arises in physical situations where events are happening randomly within some inteval in time or space.)

Theorem 38. *Suppose $Y \sim$ Poisson(λ) on interval (a, b) and it is known that $Y = 1$. Let T be the random location of the single event in (a, b). Then,*

$$P(T = t \mid Y = 1) = \frac{1}{b-a}$$

In other words, $T \mid Y \sim$ Uniform(a, b).

Proof. Suppose $t \in (a, b)$ is a possible time of the event.

$$\begin{aligned} P(T \le t \mid Y = 1) &= \frac{P(T \le t \cap Y = 1)}{P(Y=1)} && \text{definition of conditional probability} \\ &= \frac{P(\text{1 event in (a,t]} \cap \text{0 events in (t,b)})}{P(Y=1)} && \text{in other words} \end{aligned}$$

Now, consider events occurring in interval $(a, t]$. These events follow a Poisson process with mean $\lambda_1 = \lambda \times \frac{t-a}{b-a}$ by transformation of rate from original interval (a, b) to smaller interval $(a, t]$. Let the number of events in this interval be $Z_1 \sim \text{Poisson}(\lambda_1)$. Also, consider events occurring in interval (t, b). These events follow a Poisson process with mean $\lambda_2 = \lambda \times \frac{b-t}{b-a}$. Let the number of events in this interval be $Z_2 \sim \text{Poisson}(\lambda_2)$. Furthermore, the events in each of these intervals are independent because the intervals do not overlap, and by independence of events in the Poisson process, so Z_1 and Z_2 are independent. Continuing our derivation,

$$\begin{aligned} P(T \le t \mid Y = 1) &= \frac{P(\text{1 event in (a,t]} \cap \text{0 events in (t,b)})}{P(Y=1)} && \text{from above} \\ &= \frac{P(Z_1=1 \cap Z_2=0)}{P(Y=1)} && \text{new definitions} \\ &= \frac{P(Z_1=1)P(Z_2=0)}{P(Y=1)} && \text{independence} \\ &= \frac{e^{-\lambda_1}\lambda_1 e^{-\lambda_2}}{e^{-\lambda}\lambda} && \text{Poisson pmf} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}\lambda_1}{e^{-\lambda}\lambda} \\ &= \frac{\lambda_1}{\lambda} \\ &= \frac{t-a}{b-a} \end{aligned}$$

which is the cdf of Uniform(a, b). □

13.2 Expectation & Variance

Theorem 39. *If $Y \sim$ Uniform(θ_1, θ_2), then*

$$E[Y] = \frac{\theta_1 + \theta_2}{2} \quad V(Y) = \frac{(\theta_2 - \theta_1)^2}{12}$$

Proof. You should be able to derive this proof. □

13.3 Uniform in R

Function	Arguments	What it Computes
`dunif(x, min=0, max=1)`	min= θ_1, max= θ_2	$f(x)$
`punif(q, min=0, max=1)`	min= θ_1, max= θ_2	$F(q)$
`qunif(p, min=0, max=1)`	min= θ_1, max= θ_2	θ_p
`runif(n, min=0, max=1)`	min= θ_1, max= θ_2	$Y \sim$ Uniform(θ_1, θ_2)

If only the first argument is provided, then the *standard uniform* random variable is used.

13.4 Examples

Insert your own examples here learned in class. I covered one review example.

Example:

> Trucks haul concrete to a construction site with Uniform$(50, 70)$ cycle time, measured in minutes. What is the probability that the cycle time exceeds 65 minutes given that it is known to exceed 55 minutes (for example, say you have been waiting 55 minutes and want to know the

probability you will have to wait at least 10 more minutes). What are the mean and variance of the cycle time?

Let Y be the cycle time, then $Y \sim \text{Uniform}(50, 70)$. Note, $F(y) = \frac{y-50}{20}$. We seek

$$P(Y > 65 \mid Y > 55) = \frac{P(Y > 65 \cap Y > 55)}{P(Y > 55)} = \frac{P(Y > 65)}{P(Y > 55)} = \frac{1 - F(65)}{1 - F(55)} = \frac{1 - \frac{65-50}{20}}{1 - \frac{55-50}{20}} = \frac{1}{3}$$

Also,

$$E[Y] = \frac{50 + 70}{2} = 60$$

and

$$V(Y) = \frac{(70 - 50)^2}{12} = \frac{100}{3}$$

14 Normal Distribution

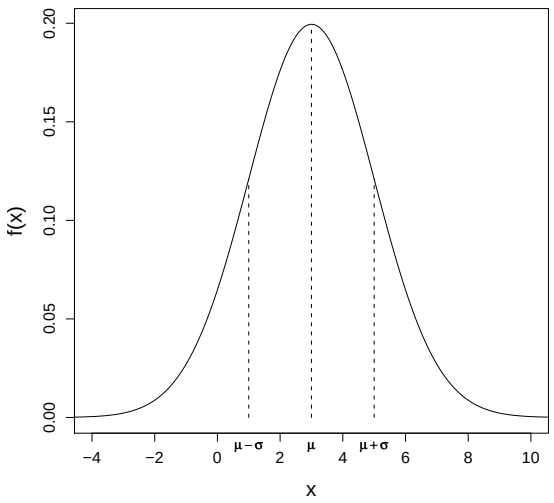
14.1 Probability Density Function

Definition: *normal random variable*

We say continuous random variable $Y \sim \text{Normal}(\mu, \sigma^2)$ or $Y \sim \text{N}(\mu, \sigma^2)$, $\sigma > 0$, $-\infty < \mu < \infty$, has normal distribution if and only if its probability density function is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < \infty$$

The following plot shows the normal pdf for $\mu = 3$ and $\sigma^2 = 4$.



Properties:

- Unimodal: there is a single peak at $y = \mu$.
- Symmetric around μ: $f(y + \mu) = f(-y + \mu)$ for all y.
- No closed-form expression for $F(y)$ exists. Numerical methods must be used.
- $\sigma > 0$ is required otherwise $f(y) < 0$, which cannot be true (because the cdf is increasing).

14.2 Expectation & Variance

Theorem 40. *Given* $Y \sim N(\mu, \sigma^2)$,

$$E[Y] = \mu \quad V(Y) = \sigma^2$$

Proof. Omitted. □

Definition: *standard normal*

> $Z \sim \mathrm{N}(0, 1)$ is said to have the standard normal distribution.

Any $Y \sim \mathrm{N}(\mu, \sigma^2)$ can be *transformed* to a standard normal random variable using the following function

$$Z = \frac{Y - \mu}{\sigma}$$

The proof is left for Stat342. This transformation is less important in today's computer age, but printed statistical tables for the normal random variable provide probabilities for Z. There are other reasons to remember this transformation; useful for finding μ and σ^2 such that the random variable satisfies certain properties.

14.3 Normal in R

Function	Arguments	What it Computes
`dnorm(x, mean=0, sd=1)`	`mean`= μ, `sd`= σ	$f(x)$
`pnorm(q, mean=0, sd=1)`	`mean`= μ, `sd`= σ	$F(q)$
`qnorm(p, mean=0, sd=1)`	`mean`= μ, `sd`= σ	ϕ_p
`rnorm(n, mean=0, sd=1)`	`mean`= μ, `sd`= σ	$Y \sim \mathrm{N}(\mu, \sigma^2)$

If only the first argument is provided, then the *standard normal* random variable is used.

14.4 Examples

You should know how to, for $Y \sim \mathrm{N}(\mu, \sigma^2)$:

- compute $P(Y > a)$, $P(a \leq Y \leq b)$, or area of shaded regions specified in graphs
- compute probability that n independent realizations of a normal random variable satisfy a particular condition, e.g. $\in [a, b]$ (just a reminder of Binomial distribution)
- find quantiles (median is interesting–to see if they get it–because of symmetry, word problems, e.g. how high in order to guarantee greater than X% of occurrences)
- $P(Z^2 < a)$ or $P(Z^2 > b)$, requires a little thinking
- Find μ, given σ such that a certain probability statement is met. Vice versa. Another use of Z.

Example:

> Suppose a soft-drink machine discharges an average μ oz. per cup. If the amount dispensed is normally distributed with standard deviation 0.3, when μ will result in overlow only 1% of the time?
>
> Let X be the amount dispensed by the machine. We are given $X \sim \mathrm{N}(\mu, 0.3^2)$.

When asked to compute μ or σ such that certain properties are satisfied, it is useful to work with the standardized value Z. We seek μ such that

$$\begin{aligned} P(X > 8) &= 0.01 \\ P(X - \mu > 8 - \mu) &= 0.01 \\ P\left(\frac{X-\mu}{0.3} > \frac{8-\mu}{0.3}\right) &= 0.01 \\ P\left(Z > \frac{8-\mu}{0.3}\right) &= 0.01 \\ P\left(Z \leq \frac{8-\mu}{0.3}\right) &= 0.99 \end{aligned}$$

We know that `qnorm(0.99)` is the quantile $\phi_{0.99}$ such that $P(Z \leq \phi_{0.99}) = 0.99$. Thus

$$\phi_{0.99} = \frac{8-\mu}{0.3} \quad \Rightarrow \quad \mu = 8 - 0.3\phi_{0.99} \approx 7.302$$

Some other quantities we might need to compute, where we set `mu`$= 7.302$.

1. $P(X > 7.5) = 1 - P(X \leq 7.5)$ is computed as `1-pnorm(7.5,mean=mu, sd=0.3)` $= 0.255$.
2. $P(7.5 \leq X \leq 8) = P(X \leq 8) - P(X \leq 7.5) =$ `pnorm(8, mean=mu, sd=0.3) - pnorm(7.5, mean=mu, sd=0.3)` $= 0.245$
3. What is the probability that 100 cups do not overflow? This is just a binomial probability $\binom{100}{0}0.1^0 0.9^{100} =$ `dbinom(0, size=100, prob=0.1)` $= 2.65614e-05$.
4. If you are given μ and asked to compute σ, make the same calculations but obtain a formula involving σ, not μ.
5. $P(X^2 < a) = P(-\sqrt{q} < X < \sqrt{a}) = P(X \leq \sqrt{a}) - P(X \leq -\sqrt{a})$ and compute as in item 2. Similarly, $P(X^2 > a) = P(X > \sqrt{a}) + P(X < -\sqrt{a})$.

15 Gamma Distribution

15.1 Probability Density Function

Introduction

We need a library of continuous random variables to characterize the distributions of continuous random outcomes visible in everyday life. So far, we have defined the Uniform and Normal distributions, but they have limited applicability in certain situations. Consider the random variable describing the time it takes for some random event to happen (e.g. bus to arrive, service at the check-out to finish, Kentucky Derby winner to finish the race, etc.). It is not reasonable to hypothesize a uniform distribution for this "wait time" because there is not concrete lower and upper limit; in addition, we expect the distribution to be notably peaked at the most likely times. Nor is it reasonable to use a normal distribution, because the pdf puts positive, albeit possibly very small, probability on negative numbers, and wait times can never be negative.

We need another kind of distribution useful for random variables with range in the positive real line. Gamma and its special cases, exponential and chi-square, come to the rescue.

Gamma pdf

Definition: *Gamma distribution*

The random variable $Y \sim \text{Gamma}(\alpha, \beta)$ is said to have a Gamma distribution with *shape parameter* α and *scale parameter* β if and only if it has probability density function

$$f(y) = \begin{cases} \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)} & 0 \leq y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Definition: *Gamma function*

The special function, appearing in the denominator of the Gamma pdf, is called the *Gamma function*. It is a definite integral, which cannot be solved analytically in the general case:

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}dy$$

Properties of the Gamma Function

The Gamma function has some special properties.

1. $\Gamma(1) = 1$

   Proof.

   $$\Gamma(1) = \int_0^{\infty} e^{-y}dy = -e^{-y}\Big|_0^{\infty} = 0 - (-1) = 1$$

   □

2. $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$

   Proof.

   $$\begin{array}{rcll} \Gamma(\alpha) & = & \int_0^{\infty} y^{\alpha-1}e^{-y}dy & \text{definition} \\ & = & -y^{\alpha-1}e^{-y}\Big|_0^{\infty} + (\alpha-1)\int_0^{\infty} y^{\alpha-2}e^{-y}dy & \text{integration by parts} \\ & = & (\alpha-1)\Gamma(\alpha-1) & \text{definition} \end{array}$$

   Notice $y^{\alpha-1}e^{-y}\big|_{y=\infty} = 0$ by L'Hospital's rule even for $\alpha > 1$, *i.e.* e^{-y} goes to 0 faster than $y^{1-\alpha}$ as $y \to \infty$. □

3. $\Gamma(n) = (n-1)!$ for any positive integer n.

   Proof.

   $$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) = (n-1)(n-2)\cdots 2\Gamma(1) = (n-1)!$$

   □

15.2 Expectation & Variance

Properties of Gamma Distribution

1. If $Y \sim \text{Gamma}(\alpha, \beta)$ and $X \sim \text{Gamma}(\alpha + 1, \beta)$, then

   $$f_X(x) = \frac{x}{\alpha\beta} f_Y(x)$$

and in general if $Z \sim \text{Gamma}(\alpha + n, \beta)$ for positive integer n, then

$$f_Z(z) = \frac{z^n}{\beta^n \alpha(\alpha+1)\cdots(\alpha+n-1)} f_Y(z) = \frac{z^n \Gamma(\alpha)}{\beta^n \Gamma(\alpha+n)} f_Y(z)$$

Thus, we can obtain the pdf of $\text{Gamma}(\alpha + n, \beta)$ easily from $\text{Gamma}(\alpha, \beta)$.

$$\begin{array}{rcll} f_X(x) & = & \frac{x^\alpha e^{-x/\beta}}{\beta^{\alpha+1}\Gamma(\alpha+1)} & \text{definition of pdf} \\ & = & \frac{x x^{\alpha-1} e^{-x/\beta}}{\beta\beta^\alpha \alpha\Gamma(\alpha)} & \text{rearrangement and using Gamma fxn properties} \\ & = & \frac{x}{\beta\alpha} f_Y(x) & \text{definition of Gamma pdf} \end{array}$$

Repeated application of the above n times, yields the general result.

2.

Theorem 41. *If $Y \sim$ Gamma(α, β), then $E[Y] = \alpha\beta$ and $V(Y) = \alpha\beta^2$.*

Proof.

$$E[Y] = \int_0^\infty y f_Y(y) dy = \alpha\beta \int_0^\infty \frac{y}{\alpha\beta} f_Y(y) dy = \alpha\beta$$

because the integral integrates the pdf of $Z \sim \text{Gamma}(\alpha + 1, \beta)$ over its entire range and must be 1. Similarly $E[Y^2] = \alpha(\alpha+1)\beta^2$, so $V(Y) = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2$. □

Other moments $E[Y^n]$ can be found with similar ease.

15.3 Gamma in R

Now, suppose $X \sim \text{Gamma}(\alpha, \beta)$ and we let `alpha`$= \alpha$ and `beta`$= \beta$, then the following R functions are useful.

Command	Computes/Generates
`dgamma(x, shape=alpha, scale=beta)`	$f(x)$
`pgamma(q, shape=alpha, scale=beta)`	$P(X \leq x)$
`qgamma(p, shape=alpha, scale=beta)`	ϕ_p such that $P(X \leq \phi_p) = p$
`rgamma(n, shape=alpha, scale=beta)`	Independent $X_1, \ldots, X_n \sim \text{Gamma}(\alpha, \beta)$.

15.4 Related Distributions

15.4.1 Chi-Square Distribution

Definition: *chi-square distribution with ν degrees of freedom*

> $Y \sim \chi^2_\nu$ is said to have a chi-square distribution with ν degrees of freedom if and only if $Y \sim \text{Gamma}(\nu/2, 2)$.

The chi-square distribution has one parameter ν, which is called the *degrees of freedom*. The main value of the chi-square distribution if for doing statistical inference, so if you take a class in statistics, you will see the chi-square distribution appear frequently.

15.4.2 Exponential Distribution

Definition: *exponential distribution*

$Y \sim$ Exponential(β), $\beta > 0$ is said to have an exponential distribution if and only if $Y \sim$ Gamma$(1, \beta)$. In other words, if and only if it has pdf

$$f(y) = \begin{cases} \frac{e^{-y/\beta}}{\beta} & 0 \leq y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Notice, that the cdf can be computed for an exponential distribution. The exponential also has some other properties of interest.

1. $F(y) = 1 - e^{-y/\beta}$

   Proof.

   $$F(y) = \int_0^y \frac{e^{-t/\beta}}{\beta} dt = -e^{-t/\beta}\Big|_0^y = -e^{-y/\beta} - (-1)$$

   □

   Notice, this implies $P(Y \leq y) = 1 - e^{-y/\beta}$ and $P(Y > y) = e^{-y/\beta}$.

2. If $Y \sim$ Exp(β), then $E[Y] = \beta$ and $V(Y) = \beta^2$.

3. The exponential distribution is *memoryless*. In english, this means that the past does not influence the future. In mathematical terms, this means $P(Y > a + b \mid Y > a) = P(Y > b)$.

   $$\begin{array}{rcll} P(Y > a+b \mid Y > a) & = & \frac{P(Y>a+b)}{P(Y>a)} & \text{definition of conditional distribution} \\ & = & \frac{e^{-(a+b)/\beta}}{e^{-a/\beta}} & \text{because } P(Y > y) = e^{-y/\beta} \\ & = & e^{-b/\beta} & \text{properties of } e \\ & = & P(Y > b) & \text{definition of } P(Y > y) \end{array}$$

15.5 Examples

1. Suppose the magnitude of an earthquake on the Richter scale follows and Exponential distribution with mean 2.4. What is the (a) probability that an earthquake exceeds 3.0? and (b) the probability the earthquake is between 2.0 and 3.0?

   Let X be the magnitude of the earthquake on the Richter scale. We are told $X \sim$ Exponential(2.4) because $E[X] = \beta = 2.4$.

   (a) $P(X > 3.0) = e^{-3.0/2.4} \approx 0.29$

   (b) $P(2.0 < X < 3.0) = P(X < 3.0) - P(X \leq 2.0) = 1 - e^{-3.0/2.4} - \left(1 - e^{2.0/2.4}\right) \approx 0.148$

2. If $Y \sim$ Exponential(β) with $P(Y > 2) = 0.0821$, what is β?

   We are given that $e^{-2/\beta} = 0.0821$, so $-2/\beta = \ln 0.0821$ and $\beta = \frac{-2}{\ln 0.0821} = 0.8$.

3. A water pumping station notices demand has exponential distribution with mean 100cfs (cubic feet per second). What capacity should the station maintain such that demand exceeds capacity with only 0.01 probability?

   Let Y be the demand and C the constant capacity the pumping station will deliver. We are told $Y \sim$ Exponential(100) and we seek C such that

   $$P(Y > C) = 0.01$$

   but then we require $e^{-C/100} = 0.01$, so $C \approx 460.517$.

16 Beta Distribution

16.1 Probability Density Function

Definition: *Beta distribution*

$Y \sim \text{Beta}(\alpha, \beta), \alpha, \beta > 0$ is said to have a Beta distribution if and only if it has pdf

$$f(y) = \begin{cases} \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)} & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Definition: *Beta function*

The function appearing in the denominator of the Beta pdf is called the *beta function* and is defined as

$$B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Proof. The reduction of the Beta function to Gamma functions is obtained using integration by parts. Let $u = y^{\alpha-1}$ and $dv = (1-y)^{\beta-1}dy$, then $du = (\alpha-1)y^{\alpha-2}dy$ and $v = \frac{-(1-y)^{\beta}}{\beta}$, so

$$\begin{aligned} B(\alpha,\beta) &= y^{\alpha-1}\left(\frac{-(1-y)^{\beta}}{\beta}\right)\Big|_0^1 + \frac{\alpha-1}{\beta}\int_0^1 y^{\alpha-2}(1-y)^{\beta+1}dy \\ &= \frac{\alpha-1}{\beta}B(\alpha-1,\beta+1) \\ &= \frac{(\alpha-1)(\alpha-2)\cdots 1}{\beta(\beta+1)\cdots(\beta+\alpha-2)}B(1,\beta+\alpha-1) \\ &= \frac{(\alpha-1)(\alpha-2)\cdots 1}{\beta(\beta+1)\cdots(\beta+\alpha-2)}\int_0^1 (1-y)^{\alpha+\beta-2}dy \\ &= \frac{(\alpha-1)(\alpha-2)\cdots 1}{\beta(\beta+1)\cdots(\beta+\alpha-2)}\left[\frac{-(1-y)^{\alpha+\beta-1}}{\alpha+\beta-1}\right]\Big|_0^1 \\ &= \frac{(\alpha-1)(\alpha-2)\cdots 1}{\beta(\beta+1)\cdots(\beta+\alpha-2)(\beta+\alpha-1)} \\ &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \end{aligned}$$

□

The following plots show various beta distributions for $\alpha = \beta$ in Fig. 6(a) and $\alpha > \beta$ in Fig. 6(b).

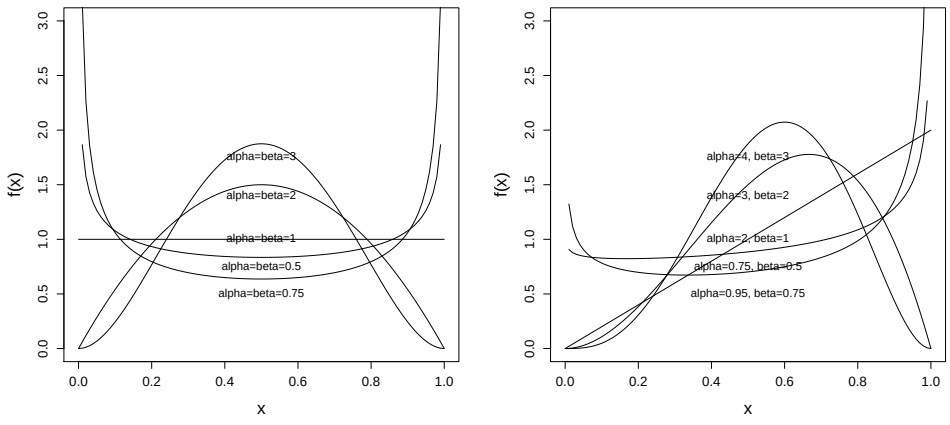
Properties

1. The standard $Uniform(0,1)$ is a special case of the Beta. Specifically $\alpha = \beta = 1$ yields

$$f(y) = \frac{y^0(1-y)^0}{B(1,1)} = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} = \frac{(2-1)!}{1} = 1$$

In general, we can use a Beta distribution to put a non-uniform distribution on any finite interval in the real line. For example, if $\theta_1 < X < \theta_2$, then define

$$Y = \frac{X-\theta_1}{\theta_2-\theta_1}$$



(a) (b)

Figure 6: Examples of Beta distributions

and observe that $0 < Y < 1$. Then, it is legitimate for $Y \sim \text{Beta}(\alpha, \beta)$. Thus, we can scale and translate X and to get Y, and then we can place a Beta distribution on Y. To obtain the distribution implied for X requires additional knowledge. For now just understand that the Beta allows us to define flexible distributions on any finite interval.

2. The CDF

$$F(y) = \int_0^y \frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha,\beta)} dt$$

is not analytically available, in general. However, for α and β integers, we have

$$P(Y \leq y) = P(X \geq a)$$

where $Y \sim \text{Beta}(\alpha, \beta)$ and $X \sim \text{Binomial}(\alpha + \beta - 1, y)$.

3.

Theorem 42. *If $Y \sim$ Beta(α, β), then*

$$E[Y] = \frac{\alpha}{\alpha+\beta} \qquad V(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Proof. This proof also demonstrates computation of $E[Y^n]$ for arbitrary integer powers n.

$$\begin{aligned}
E[Y] &= \int_0^1 \frac{y y^{\alpha-1}(1-y)^{\beta-1}}{B}(\alpha,\beta)dy \\
&= \frac{1}{B(\alpha,\beta)}\int_0^1 y^{\alpha}(1-y)^{\beta-1}dy \\
&= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \\
&= \frac{\Gamma(\alpha+1)\Gamma(\beta)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+1)} \\
&= \frac{\alpha}{\alpha+\beta} \\
E[Y^2] &= \frac{1}{B(\alpha,\beta)}\int_0^1 y^{\alpha+1}(1-y)^{\beta-1}dy \\
&= \frac{B(\alpha+2,\beta)}{B(\alpha,\beta)} \\
&= \frac{(\alpha+1)\alpha}{(\alpha+\beta)(\alpha+\beta+1)} \\
&\vdots \\
E[Y^n] &= \frac{(\alpha+n-1)(\alpha+n-2)\cdots(\alpha+1)\alpha}{(\alpha+\beta)(\alpha+\beta+1)\cdots(\alpha+\beta+n-1)} \\
&= \frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+n)}
\end{aligned}$$

For variance,

$$\begin{aligned}
V(Y) &= E[Y^2] - (E[Y])^2 \\
&= \frac{(\alpha+1)\alpha}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{(\alpha+1)\alpha(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}
\end{aligned}$$

□

4. What is the mode of a Beta distribution?

   We do the derivation for $\alpha, \beta > 1$. The mode occurs where the distribution reaches a maximum, i.e. where the derivative is 0.

$$\begin{aligned}
\frac{df(y)}{dy} &= 0 \\
\frac{(\alpha-1)y^{\alpha-2}(1-y)^{\beta-1}}{B(\alpha,\beta)} - \frac{(\beta-1)y^{\alpha-1}(1-y)^{\beta-2}}{B(\alpha,\beta)} &= 0 \\
(\alpha-1)y^{\alpha-2}(1-y)^{\beta-1} - (\beta-1)y^{\alpha-1}(1-y)^{\beta-2} &= 0 \\
(\alpha-1)(1-y) - (\beta-1)y &= 0 \\
y &= \frac{\alpha-1}{\alpha+\beta-2}
\end{aligned}$$

16.2 Beta Distribution in R

Suppose $X \sim$ Beta(α, β), and let `alpha`$= \alpha$ and `beta`$= \beta$.

R Command	What it Computes
`dbeta(x, shape1=alpha, shape2=beta)`	$f(x)$
`pbeta(q, shape1=alpha, shape2=beta)`	$P(X \leq q)$
`qbeta(p, shape1=alpha, shape2=beta)`	Find ϕ_p such that $P(X \leq \phi_p) = p$
`rbeta(n, shape1=alpha, shape2=beta)`	Generate independent $X_1, X_2, \ldots, X_n \sim$ Beta(α, β)

16.3 Examples

1. Suppose a random variable has pdf

$$f(y) = \begin{cases} Cy^3(1-y)^2 & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   What is the distribution and what is C?

   The random variable has a Beta distribution with $\alpha = 4$ and $\beta = 3$. The constant is obviously $C = \frac{1}{B(4,3)} = \frac{\Gamma(7)}{\Gamma(3)\Gamma(4)} = 60$.

2. Suppose the above pdf defines the random probability that a person will buy a product after they have seen an advertisement for it. What is the probability this probability will exceed 0.90?

   Let $Y \sim$ Beta$(4, 3)$ be the probability a person buys the product. We seek $P(Y > 0.9) = 1 -$ `pbeta(0.9, shape1=4, shape2=3)` $= 0.01585$.

3. What is $E[Y]$ and $V(Y)$?

   $E[Y] = \frac{4}{4+3} \approx 0.57$. $V(Y) = \frac{4 \times 3}{7^2 \times 8} \approx 0.0306$.

4. How could you use `R` to verify the above calculations if you suddenly doubted your cheatsheet?

```
> y <- rbeta(1000, shape1=4, shape2=3)
> sum(y>0.9)/1000
[1] 0.018
> mean(y)
[1] 0.576
> var(y)
[1] 0.0323
```

5. Find $F(y)$.

$$\begin{aligned} F(y) &= \int_0^y 60t^3(1-t)^2 dt \\ &= 60 \int_0^y (t^3 - 2t^4 + t^5) dt \\ &= 60 \left[\frac{t^4}{4} - \frac{2t^5}{5} + \frac{t^6}{6} \right] \Big|_0^y \\ &= 60 \left[\frac{y^4}{4} - \frac{2y^5}{5} + \frac{y^6}{6} \right] \\ &= 0.01585 \quad \text{verifies calculation in 2} \end{aligned}$$

6. Suppose the proportion of downtime of a assembly line Y has Beta distribution with $\alpha = 1$ and $\beta = 2$. The cost due to downtime is $C = 10 + 20Y + 4Y^2$. What is $E[C]$ and $V(C)$? [Note, I have not verified the calculations; the approach is correct.]

   We have $Y \sim \text{Beta}(1, 2)$, and $E[Y] = \frac{1}{3}$. Also, we know $E[Y^2] = \frac{2\times1}{3\times4} = \frac{1}{6}$, $E[Y^3] = \frac{3\times2\times1}{3\times4\times5} = \frac{1}{10}$, and $E[Y^4] = \frac{4\times3\times2\times1}{3\times4\times5\times6} = \frac{1}{15}$. We'll use these below.

$$E[C] = 10 + 20E[Y] + 4E[Y^2] = 10 + \frac{20}{3} + \frac{4}{6} = \frac{52}{3}$$

$$\begin{aligned}
V(C) &= E[C^2] - (E[C])^2 \\
E[C^2] &= E[(10 + 20Y + 4Y^2)^2] \\
&= E[100 + 400Y + 40Y^2 + 400Y^2 + 80Y^3 + 16Y^4] \\
&= 100 + 400E[Y] + 440E[Y^2] + 80E[Y^3] + 16E[Y^4] \\
&= 100 + \frac{400}{3} + \frac{440}{6} + \frac{80}{10} + \frac{16}{15} \\
&= \frac{4736}{15} \\
V(C) &= \frac{4736}{15} - \left(\frac{52}{3}\right)^2 \\
&\approx 15.289
\end{aligned}$$

Part IV

Moment Generating Functions

17 Discrete Random Variables

17.1 Definitions

Moments

Definition: *kth moment about the origin*

The *kth moment about the origin* is $E[Y^k]$.

Definition: *kth central moment*

The *kth central moment* is $E[(Y - \mu)^k]$, where $\mu = E[Y]$ is the first moment about the origin.

Moment-Generating Function

Definition: *moment-generating function (mgf)*

For a random variable Y, the moment-generating function

$$m(t) = E\left[e^{tY}\right]$$

is defined if there exists $b > 0$ such that $m(t) < \infty$ for $|t| \leq b$.

Properties:

1. $m(0) = E\left[e^0\right] = 1$
2. The sum $\sum_y p(y) = 1$ of course converges, but $m(t) = \sum_y e^{ty} p(y)$ may not converge because $e^{ty} > 1$ whenever $ty > 0$.
3. Where do *moments* come in?

   In what follows, we assume $|t| < b$ such that the sum converges ($m(t)$ exists).

   $$\begin{array}{rcll} m(t) & = & \sum_y e^{ty} p(y) & \text{definition of mgf} \\ & = & \sum_y \sum_i \left(\frac{(ty)^i}{i!}\right) p(y) & \text{Taylor expansion of } e^{ty} \\ & = & \sum_k \sum_y \frac{(ty)^k}{k!} p(y) & \text{exchange of sums OK because of convergence} \\ & = & \sum_k \frac{t^k}{k!} E[Y^i] & \text{definition of moments around origin} \end{array}$$

   Thus, $m(t)$ can be viewed as a function of all of the random variable's moments about the origin.
4. And in fact, the moment-generating function can also *generate* moments, just as the name suggests.

   Theorem 43. *If $m(t)$ exists, then*

   $$\left.\frac{d^k m(t)}{dt^k}\right|_{t=0} = m^{(k)}(0) = E[Y^k]$$

   Proof.

   $$\begin{array}{rcll} m'(t) & = & \frac{d}{dt} \sum_y e^{ty} p(y) & \\ & = & \sum_y \frac{de^{ty}}{dt} p(y) & \text{convergence allows exchange of derivative and sum} \\ & = & \sum_y y e^{ty} p(y) & \\ m'(0) & = & \sum_y y p(y) & \\ & = & E[Y] & \end{array}$$

   Since $\frac{d^k e^{ty}}{dt^k} = y^k e^{ty}$, it is easy to see this result generalizes to the kth derivative. □
5. We will not prove this result, but when the $m(t)$ exists, it is unique. This implies that if two random variables X and Y have the same mgf for all $|t| < b$ and some $b > 0$, then X and Y must have the same distribution (i.e. they have the same pmf). It also implies that if you know Y has an mgf $m(t)$ and you recognize that mgf, you also know the distribution of Y.
6. We have already seen

   $$\begin{array}{rcl} E[aY + b] & = & aE[Y] + b \\ V(aY + b) & = & a^2 V(Y) \end{array}$$

   and this is a result that is perhaps more easily obtained by mgf's.

   Proof. Suppose Y has mgf $m_Y(t)$ and $X = g(Y) = aY + b$. Then,

   $$m_X(t) = E\left[e^{tg(Y)}\right] = E\left[e^{t(aY+b)}\right] = E\left[e^{tb} e^{atY}\right] = e^{tb} m_Y(at)$$

   With moment generating function in hand,

   $$E[X] = m'(0) = be^{tb} m_Y(at) + ae^{tb} m'_Y(at)\big|_{t=0} = b + aE[Y]$$

   and

   $$E[X^2] = m''(0) = b^2 e^{tb} m_Y(at) + 2abe^{tb} m'_Y(at) + a^2 e^{tb} m''_Y(at)\big|_{t=0} = b^2 + 2abE[Y] + a^2 E[Y^2]$$

   so

   $$V(X) = b^2 + 2abE[Y] + a^2 E[Y^2] - (b + aE[Y])^2 = a^2 E[Y^2] - a^2 (E[Y])^2 = a^2 V(Y)$$

   □

We write up the result we used above as a corollary.

Corollary 44. *If Y has mgf $m_Y(t)$, then $X = aY + b$ has mgf $m_X(t) = e^{tb}m_Y(at)$.*

17.2 Moment Generating Functions

Theorem 45. *If $Y \sim$ Poisson(λ), then $m_Y(t) = e^{\lambda(e^t-1)}$.*

Proof. The idea, as always, is to manipulate the summand until a recognizable series appears.

$$\begin{aligned} m_Y(t) &= \sum_{y=0}^{\infty} e^{ty} p_Y(y) \\ &= \sum_{y=0}^{\infty} e^{ty} \frac{\lambda^y e^{-\lambda}}{y!} \\ &= \sum_{y=0}^{\infty} \frac{(e^t\lambda)^y e^{-\lambda}}{y!} \\ &= e^{-\lambda} \sum_{y=0}^{\infty} \frac{(e^t\lambda)^y}{y!} \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= e^{\lambda(e^t-1)} \end{aligned}$$

□

Theorem 46. *If $Y \sim$ Binomial(n, p), then $m_Y(t) = [e^t p + (1-p)]^n$*

Proof.

$$\begin{aligned} m_Y(t) &= \sum_{y=0}^{n} e^{ty} \tbinom{n}{y} p^y (1-p)^{n-y} && \text{pmf of Binomial} \\ &= \sum_{y=0}^{n} \tbinom{n}{y} (pe^{ty})^y (1-p)^{n-y} \\ &= [pe^{ty} + (1-p)]^n && \text{Binomial expansion} \end{aligned}$$

□

Theorem 47. *If $Y \sim$ Geometric(p), then $m_Y(t) = \frac{pe^t}{1-e^t(1-p)}$.*

Proof.

$$\begin{aligned} m_Y(t) &= \sum_{y=1}^{\infty} e^{yt} p(1-p)^{y-1} && \text{pmf of Geometric} \\ &= pe^t \sum_{y=1}^{\infty} e^{t(y-1)} (1-p)^{y-1} && \text{get matching } y-1 \text{ exponents} \\ &= pe^t \sum_{x=0}^{\infty} [e^t(1-p)]^x && \text{change of variable } x = y-1 \text{ and consolidate} \\ &= pe^t \tfrac{1}{1-(1-p)e^t} && \textstyle\sum_0^{\infty} p^x = \frac{1}{1-p} \end{aligned}$$

□

17.3 Examples

Now we demonstrate how to use the mgf's.

Example:

1. We will first demonstrate $E[Y] = V(Y) = \lambda$ for $Y \sim$ Poisson(λ). Recall $m(t) = e^{\lambda(e^t-1)}$, then

$$\begin{aligned} m'(t) &= \lambda e^t e^{\lambda(e^t-1)}\Big|_{t=0} = \lambda \\ m''(t) &= \lambda e^t e^{\lambda(e^t-1)} + \lambda^2 e^{2t} e^{\lambda(e^t-1)}\Big|_{t=0} \\ &= \lambda + \lambda^2 \\ V(Y) &= \lambda + \lambda^2 - \lambda^2 = \lambda \end{aligned}$$

2. We will also use the mgf to compute the mean and variance of the Binomial random variable.

$$\begin{aligned} m'(t) &= n\left[e^t p + (1-p)\right]^{n-1} e^t p\Big|_{t=0} = n\left[p + 1 - p\right]^{n-1} p = np \\ m''(t) &= n\left[e^t p + (1-p)\right]^{n-1} e^t p + n(n-1)(e^t p)^2 \left[e^t p + (1-p)\right]^{n-2}\Big|_{t=0} = np + n(n-1)p^2 \\ V(Y) &= np + n(n-1)p^2 - (np)^2 = np - np^2 = np(1-p) \end{aligned}$$

3. Suppose you know $m_Y(t) = e^{3.2(e^t-1)}$, then what is the distribution of Y? You recognize $m_Y(t)$ has the form of a Poisson random variable and $\lambda = 3.2$, so $Y \sim$ Poisson(3.2).
4. Suppose $m_Y(t) = (0.37e^t + 0.63)^{10}$, then $Y \sim$ Binomial$(10, 0.37)$.
5. Suppose $m_X(t) = \frac{e^t}{6} + \frac{2e^{2t}}{6} + \frac{3e^{3t}}{6}$, then what is the distribution of X?
We return to the definition $m(t) = \sum_y e^{tx} p_X(x)$ and see that X is an un-named random variable with range $X(S) = \{1, 2, 3\}$ and pmf

$$p_X(1) = \frac{1}{6} \quad p_X(2) = \frac{1}{3} \quad p_X(3) = \frac{1}{2}$$

18 Continuous Random Variable

18.1 Definitions

All definitions are as before, but of course sums are replaced with integrals.

$$\begin{aligned} E[Y^k] &= \int_{-\infty}^{\infty} y^k f(y)dy && k\text{th moment around the origin} \\ E[(Y-\mu)^k] &= \int_{-\infty}^{\infty} (y-\mu)^k f(y)dy && k\text{th central moment} \\ m(t) &= E[e^{tY}] = \int_{-\infty}^{\infty} e^{ty} f(y)dy && \text{moment-generating function} \end{aligned}$$

The properties of the mgf are little unchanged too Properties:

1. $m(t)$ is defined if $m(t) < \infty$ for all $|t| \le b$ for some $b > 0$.
2. $\frac{d^k m(t)}{dt^k}\Big|_{t=0} = E[Y^k]$. We can prove this result by moving the derivative inside the integral in the definition and using the definition of moments around the origin.
3. $m(t)$ is unique
4. $m_{aY+b}(t) = e^{tb} m_Y(at)$ gives the mgf of linear function of random variables with known mgf's.

18.2 Moment-Generating Functions

Theorem 48. *If $Y \sim$ Gamma(α, β), then $m_Y(t) = (1-\beta t)^{-\alpha}$.*

Proof. These proofs proceed by manipulating the integrand until a pdf appears, so the integral becomes 1.

$$\begin{aligned} m_Y(t) &= \int_{-\infty}^{\infty} e^{ty} \frac{y^{\alpha-1} e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)} dy && \text{definition of Gamma pdf} \\ &= \frac{1}{\beta^\alpha} \int_{-\infty}^{\infty} \frac{y^{\alpha-1} e^{-y\left(\frac{1}{\beta} - t\right)}}{\Gamma(\alpha)} dy \\ &= \frac{\left(\frac{\beta}{1+\beta t}\right)^\alpha}{\beta^\alpha} \int_{-\infty}^{\infty} \frac{y^{\alpha-1} e^{-y\left(\frac{1}{\beta} - t\right)}}{\Gamma(\alpha)\left(\frac{\beta}{1+\beta t}\right)^\alpha} dy \\ &= \frac{\left(\frac{\beta}{1+\beta t}\right)^\alpha}{\beta^\alpha} && \text{integrand is pdf of Gamma}(\alpha, \tfrac{\beta}{1+\beta t}) \\ &= \frac{1}{(1+\beta t)^\alpha} \end{aligned}$$

□

Theorem 49. *If $Y \sim$ Normal(μ, σ^2), then $m_Y(t) = e^{\mu t + \sigma^2 t^2/2}$.*

Proof.

$$\begin{aligned}
m_Y(t) &= \int_{-\infty}^{\infty} e^{yt} \tfrac{1}{\sqrt{2\pi}\sigma} \exp\left[-\tfrac{(y-\mu)^2}{2\sigma^2}\right] dy && \text{definition of Normal pdf} \\
&= \tfrac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\tfrac{(y-\mu)^2}{2\sigma^2} + yt\right] dy \\
&= \tfrac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\tfrac{[y-(\sigma^2 t+\mu)]^2}{2\sigma^2} + \tfrac{\sigma^2 t^2}{2} + \mu t\right] dy && \text{complete the square} \\
&= \tfrac{e^{\frac{\sigma^2 t^2}{2}+\mu t}}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\tfrac{[y-(\sigma^2 t+\mu)]^2}{2\sigma^2}\right] dy \\
&= e^{\frac{\sigma^2 t^2}{2}+\mu t} && \text{integrand is pdf of Normal}(\sigma^2 t + \mu, \sigma^2)
\end{aligned}$$

□

Theorem 50. *If $Y \sim$ Uniform(a, b), then $m_Y(t) = \frac{e^{ta} - e^{tb}}{t(b-a)}$.*

Proof.

$$\begin{aligned}
m_Y(t) &= \int_a^b \tfrac{e^{yt}}{b-a} dy && \text{pdf of Uniform} \\
&= \tfrac{1}{t(b-a)} e^{yt}\Big|_a^b \\
&= \tfrac{e^{bt} - e^{at}}{t(b-a)}
\end{aligned}$$

□

18.3 Examples

1. Suppose $Y \sim$ Normal(μ, σ^2). We will use the linear function result to derive the distribution of $X = Y - \mu$ and $Z = \frac{Y-\mu}{\sigma}$.

   $$m_X(t) = e^{-\mu t} m_Y(t) = e^{-\mu t} e^{\mu t + \sigma^2 t^2/2} = e^{\sigma^2 t^2/2}$$

   By uniqueness of mgf, we conclude $X \sim$ Normal$(0, \sigma^2)$. This is intuitively obvious, because we have merely shifted all values of the random variable left by μ. However, not all distributions, when shifted, give the same distribution with different parameters. For example, for $Y \sim$ Gamma(α, β), $Y - \alpha\beta$ does *not* follow a Gamma distribution.

   $$m_Z(t) = e^{-\mu t/\sigma} m_Y(t/\sigma) = e^{-\mu t/\sigma} e^{\mu t/\sigma + \sigma^2 t^2/\sigma^2/2} = e^{t^2/2}$$

   Again, by uniqueness, we conclude $Z \sim$ Normal$(0, 1)$ follows a standard normal distribution, as claimed, but not proven earlier.

2. Verify $E[Y] = \mu$ and $V(Y) = \sigma^2$ for $Y \sim$N(μ, σ^2).

   $$E[Y] = m'(0) = e^{\mu t + \sigma^2 t^2/2}(\mu + \sigma^2 t)\Big|_{t=0} = \mu$$

   and

   $$E[Y^2] = m''(0) = e^{\mu t + \sigma^2 t^2/2}(\mu + \sigma^2 t)^2 + e^{\mu t + \sigma^2 t^2/2}\sigma^2\Big|_{t=0} = \mu^2 + \sigma^2$$

   so

   $$V(Y) = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$$

3. What is the distribution of Y if $m_Y(t) = (1 - 4t)^{-2}$? $Y \sim$ Gamma$(2, 4)$.

4. What is the distribution of Y if $m_Y(t) = \frac{1}{1-3.2t}$? $Y \sim$ Gamma$(1, 3.2)$.

5. What is the distribution of Y if $m_Y(t) = e^{-5t+6t^2}$? $Y \sim$ Normal$(-5, 12)$.

Part V
Multivariate Random Variables

19 Definitions

So far, we have dealt with *univariate*, or single-valued, random variables. We observe the outcome of one random experiment and then map the result to one real number. In many circumstances, it will become necessary to map the result of the random experiment to multiple random numbers. For example,

- Suppose we observe one apple tree during the growing season and we count $Y_1 =$ the number of apples it produces and $Y_2 =$ the number of mites observed on a random subset of 20 leaves. Both Y_1 and Y_2 are discrete random variables that map the outcome of a random experiment (of a tree growing through its growing season) to the real number line.
- Suppose I collect the exam score of 23 individuals $(X_1, X_2, \ldots, X_{23})$ by giving the same exam to all individuals in the class.

In both cases, we may be interested in learning the relationships between the random numbers. These relationships are unpredictable in detail because of the randomness, but predictable in trends. For example, the count of apples Y_1 may decline as the number of mites Y_2 increases. Also, scores may tend to increase together if lectures were especially clear, or decrease together if the instructor assigned no homeworks covering the material.

In what follows, we will focus on bivariate random variables (Y_1, Y_2), but the definitions and results can extend to arbitrary n-variate random variables $(Y_1, Y_2, \ldots, Y_n)$.

Definition: *multivariate probability mass function (pmf)*

For discrete random variables (Y_1, Y_2), the *multivariate probability mass function* (pmf) is defined as

$$p(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2)$$

the joint probability of events $\{Y_1 = y_1\}$ and $\{Y_2 = y_2\}$.

Definition: *multivariate cumulative distribution function (cdf)*

For any random variables (Y_1, Y_2), the *multivariate cumulative distribution function* (cdf) is defined as

$$F(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2), \quad \text{for all } y_1, y_2 \in \Re$$

For discrete random variables, it is, specifically,

$$F(y_1, y_2) = \sum_{t_1 \leq y_1} \sum_{t_2 \leq y_2} p(t_1, t_2)$$

Definition: *joint probability density function (pdf)*

If it exists, the *joint probability density function* (pdf) is $f(t_1, t_2)$ such that

$$F(y_1, y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f(t_1, t_2) dt_2 dt_1$$

and (Y_1, Y_2) are said to be jointly continuous random variables with pdf $f(y_1, y_2)$.

19.1 Properties

The properties of the cdf are as follows

Properties:

1. $F(-\infty,-\infty) = F(-\infty,y_2) = F(y_1,-\infty) = 0$ because all of these involve the event $\{Y_1 < -\infty\}$ or $\{Y_2 < -\infty\}$ which are both the empty set, and hence have probability 0.
2. $F(\infty,\infty) = P(Y_1 < \infty, Y_2 < \infty) = 1$ because both events are always true by definition of random variable.
3. The extension of result $P(a < Y \leq b) = F(b) - F(a)$ for univariate random variables is show here. For $y_1^* \geq y_1$ and $y_2^* \geq y_2$

$$F(y_1^*,y_2^*) - F(y_1^*,y_2) - F(y_1,y_2^*) + F(y_1,y_2) \geq 0$$

Proof. The proof uses an old result $P(A \cap \overline{B}) = P(A) - P(B)$ when $B \subset A$, which can be extended to

$$P(A \cap \overline{B} \cap C) = P(A \cap C) - P(B \cap C)$$

for any event C. Then,

$$\begin{aligned} F(y_1^*,y_2^*) - F(y_1^*,y_2) - F(y_1,y_2^*) + F(y_1,y_2) &= P(Y_1 \leq y_1^*, Y_2 \leq y_2^*) - P(Y_1 \leq y_1^*, Y_2 \leq y_2) \\ &\quad - F(y_1,y_2^*) + F(y_1,y_2) \\ &= P(Y_1 \leq y_1^*, y_2 < Y_2 \leq y_2^*) - F(y_1,y_2^*) + F(y_1,y_2) \\ &= P(Y_1 \leq y_1^*, y_2 < Y_2 \leq y_2^*) - P(Y_2 \leq y_1, y_2 < Y_2 \leq y_2^*) \\ &= P(y_1 < Y_1 \leq y_1^*, y_2 < Y_2 \leq y_2^*) \geq 0 \end{aligned}$$

The proof is finished when the expression is recognized as a probability. □

The properties of the pdf are as follows

Properties:

1. $f(y_1,y_2) \geq 0$ for all $y_1, y_2 \in \Re$.
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y_1,y_2)dy_1dy_2 = 1$
3. $P(a_1 < Y_1 \leq b_1, a_2 < Y_2 \leq b_2)$ is now interpretted as the *volume* underneath the surface $f(y_1,y_2)$ sitting above the (y_1,y_2) plane where random points (Y_1,Y_2) live.
4. As before, the pdf $f(y_1,y_2)$ gives us the relative likelihood that the random variable (Y_1,Y_2) will be located close to point (y_1,y_2). Specifically, as $\Delta \to 0$, then

$$f(y_1,y_2)\Delta^2 \approx P(y_1 - \Delta/2 < Y_1 \leq y_1 + \Delta/2, y_2 - \Delta/2 < Y_2 \leq y_2 + \Delta/2)$$

19.2 Examples

19.2.1 Discrete Example

Suppose a bag is filled with 300 red beads, 200 green beads, and 50 green beads. You reach in and blindly select 10 beads with replacement. For every red bead, you earn \$1. For every blue bead, you win \$2, and for every green bead, you win \$10. Let R be the number of red beads you draw and W your winnings. The bivariate random variable (R,W) is the focus of our interest here.

The sample space is finite and can be enumerated.

$$S = \{(r,r,\ldots,r),(r,r,\ldots,r,b),(r,r,\ldots,r,g),\ldots\}$$

Each of these events has a probability that we can compute using counting methods.

The random variable R is the number of red beads, and by itself, has a binomial distribution. The random variable W is a function of the number of red, blue, and green beads. The ranges are individually

$$R(S) = \{10, 9, 9, \ldots\} \quad W(S) = \{10, 11, 19, \ldots\}$$

and together

$$\{(10, 10), (9, 11), (9, 19), \ldots\}$$

The same pair may appear multiple times in this list, but as usual, we can compute the pmf by summing over the probability of all outcomes giving the same (R, W) value. For example,

$$\begin{aligned} p(10, 10) &= \left(\frac{300}{550}\right)^{10} \\ p(9, 11) &= \binom{10}{9}\left(\frac{300}{550}\right)^{9}\left(\frac{200}{550}\right) \\ &\vdots \end{aligned}$$

We can compute the probability of any joint interval as demonstrated below

$$P(8 \leq R \leq 10, 10 \leq W \leq 15) = \sum_{r=8}^{10} \sum_{w=10}^{15} p(r, w)$$

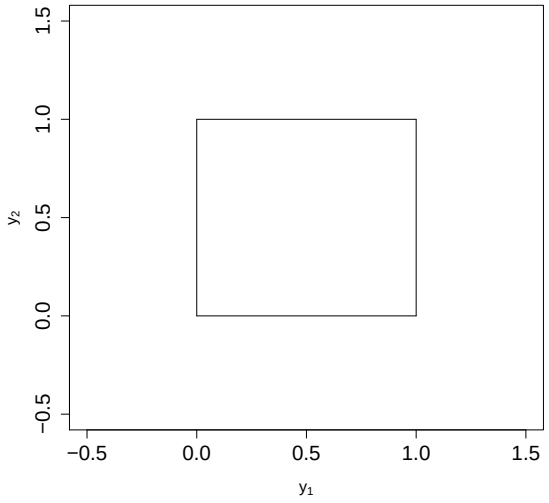
Alternatively, we can use the last property of cdf to get the result

$$P(8 \leq R \leq 10, 10 \leq W \leq 15) = F(10, 15) - F(10, 10) - F(8, 15) + F(8, 10)$$

although this will generally be more work unless the cdf $F(r, w)$ is already known.

19.2.2 Continuous Example

Bivariate Uniform Distribution



Suppose you throw a dart at the above target (square from $(0,0)$ to $(1,1)$) *without aiming* and you ignore all throws landing outside the target. The random location of the dart (Y_1, Y_2) is a bivariate random variable. By design, this random variable should be equally likely to land in all possible spots in the square, so we expect

$$f(y_1, y_2) = \begin{cases} c & 0 \le y_1, y_2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

to be some constant c on the target and 0 elsewhere.

What is c? Using the properties of the pdf, we know

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c dy_1 dy_2 &= 1 \\ \int_0^1 \left[\int_0^1 c dy_1 \right] dy_2 &= 1 \\ \int_0^1 \left[c y_1 |_0^1 \right] dy_2 &= 1 \\ \int_0^1 c dy_2 &= 1 \\ c y_2 |_0^1 = c &= 1 \end{aligned}$$

We have shown the gory details of the integration to remind you, if you have forgotten, how to deal with double integrals.

Thus, our first bivariate distribution is the bivariate uniform on the unit square

$$f(y_1, y_2) = \begin{cases} 1 & 0 \le y_1, y_2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

What is the cdf?

$$\begin{aligned} F(y_1, y_2) &= \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} dt_1 dt_2 \\ &= \int_0^{y_2} \left[\int_0^{y_1} dt_1 \right] dt_2 \\ &= \int_0^{y_1} [t_1 |_0^{y_1}] \, dt_2 \\ &= \int_0^{y_1} y_1 dt_2 \\ &= y_1 t_2 |_0^{y_1} \\ &= y_1 y_2 \end{aligned}$$

Now, you could define a uniform over any shape in the (y_1, y_2) plane, such as the circle with radius r centered at $(0,0)$. Each time the appropriate $c = 1/\text{area}$ needs to be computed.

And probabilities are computed as

$$\begin{aligned} P(0.1 \le Y_1 \le 0.2, 0.3 \le Y_2 \le 0.7) &= \int_{0.1}^{0.2} \int_{0.3}^{0.7} dy_2 dy_1 \\ &= (0.2 - 0.1) \times (0.7 - 0.3) = 0.04 \end{aligned}$$

It is usually easiest to compute these probabilities directly by integration, but the formula derived in properties also works

$$\begin{aligned} P(0.1 \le Y_1 \le 0.2, 0.3 \le Y_2 \le 0.7) &= F(0.2, 0.7) - F(0.2, 0.3) - F(0.1, 0.7) + F(0.1, 0.3) \\ &= 0.2 \times 0.7 - 0.2 \times 0.3 - 0.1 \times 0.7 + 0.1 \times 0.3 \\ &= 0.04 \end{aligned}$$

Arbitrary Distribution

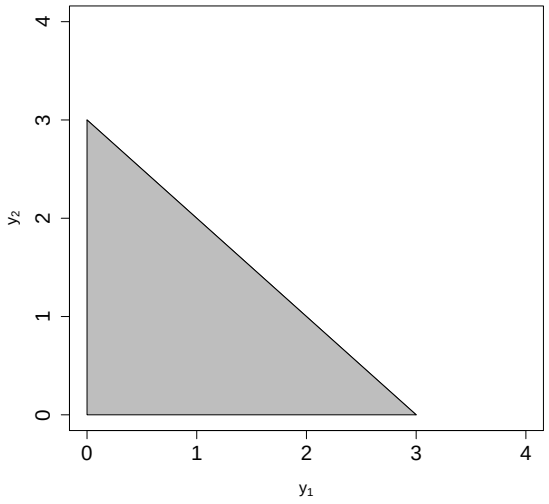
Let us consider an unnamed, arbitrary bivariate distribution

$$f(y_1, y_2) = \begin{cases} e^{-(y_1+y_2)} & 0 < y_1, y_2 \\ 0 & \text{otherwise} \end{cases}$$

Let us compute $P(Y_1 < 1, Y_2 > 5)$ and $P(Y_1 + Y_2 > 3)$. When you are computing the probability of a rectangular region (the first probability) and the allowable range of the random variables is also rectangular (this example satisfies the requirement because (Y_1, Y_2) exists on the positive quadrant), then the necessary integrals are easy to compute. If either the probability region (second probability) or the allowable range is not rectangular, then care must be taken in setting up the limits of integration. In this case, it is helpful to draw a picture of the non-rectangular range.

$$\begin{aligned} P(Y_1 < 1, Y_2 > 5) &= \int_0^1 \int_5^\infty e^{-(y_1+y_2)} dy_2 dy_1 \\ &= \int_0^1 e^{-y_1} \left[-e^{-y_2} \Big|_5^\infty \right] dy_1 \\ &= \int_0^1 e^{-y_1} e^{-5} dy_1 \\ &= e^{-5} \, -e^{-y_1} \Big|_0^1 \\ &= e^{-5} \left(1 - e^{-1}\right) = 0.00426 \end{aligned}$$

The area over which we want to integrate is defined by $y_1 + y_2 < 3$ or $y_2 < 3 - y_1$ (and of course also the range $y_1, y_2 > 0$). It is shown in the plot below. If we let y_1 range 0 to 3, then y_2 ranges from 0 up to the line $3 - y_1$. These are our limits of integration.



$$\begin{aligned}
P(Y_1+Y_2>3) &= \int_{y_1=0}^{3}\int_{y_2=0}^{3-y_1} e^{-(y_1+y_2)}dy_2dy_1 \\
&= \int_0^3 e^{-y_1}\left[-e^{-y_2}\Big|_0^{3-y_1}\right]dy_1 \\
&= \int_0^3 e^{-y_1}\left[1-e^{y_1-3}\right]dy_1 \\
&= \int_0^3 \left[e^{-y_1}-e^{-3}\right]dy_1 \\
&= \left[-e^{-y_1}-e^{-3}y_1\right]_0^3 \\
&= -e^{-3}+1-3e^{-3} = 1-4e^{-3} = 0.80085
\end{aligned}$$

Dirichlet Distribution

There is a generalization of the Beta distribution to more than one random variable. Suppose (Y_1, Y_2, Y_3) represent proportions or probabilities such that $Y_1 + Y_2 + Y_3 = 1$. Examples include

- Y_i is the proportion of ingredient i in a mixture of three ingredients.
- Y_i is the proportion of a sample that fall in category i, when there are a total of 3 categories. For example, the number of people in a sample who (1) don't exercise, (2) exercise up to twice a week, and (3) exercise more than twice a week.

Notice, that because of the constraint $Y_3 = 1-Y_1-Y_2$ is no longer random once the bivariate random variable (Y_1, Y_2) has been defined. Thus, there are really only two random variables in the collection (Y_1, Y_2, Y_3). A joint continuous pdf for (Y_1, Y_2) is given by

$$f(y_1,y_2) = \begin{cases} 360y_1^2y_2(1-y_1-y_2) & 0 \le y_1, y_2 \le 1, y_1+y_2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(**Note.** The conditions that appear on the right-hand-side are now very important as they indicate all constraints on the random variable. If you ignore them, you set your limits of integration incorrectly, and get incorrect answers.)

The above pdf is a specific example of the generalization of the Beta(α, β) distribution to two dimensions. The generalized Beta distribution is called the Dirichlet distribution (if you want to look it up, or remember it later). The random variable (Y_1, Y_2) lives on the triangle, with vertices $\{(0,0),(1,0),(0,1)\}$.

There are a number of questions I could ask you about this distribution (listed below). If you have questions about any of these, be sure to ask.

- Verify that $f(y_1, y_2)$ defines a pdf. In other words, show
$$\int_0^1\int_0^{1-y_1} 360y_1^2y_2(1-y_1-y_2)dy_2dy_1 = 1$$
- Find the constant (in this case given already as 360) such that $f(y_1, y_2)$ is a pdf.
- What is the probability that $Y_1 \le 0.5$ and $Y_2 \le 0.21$?
- What is the probability that $Y_1 + Y_2 \le 0.5$?

20 Marginal Distributions

20.1 Definitions

Definition: *marginal pmf/pdf*

If Y_1, Y_2 are discrete random variables with joint pmf $p(y_1, y_2)$, then the marginal pmf's are

$$p_{Y_1}(y_1) = \sum_{y_2} p(y_1, y_2) \quad p_{Y_2}(y_2) = \sum_{y_1} p(y_1, y_2)$$

If Y_1, Y_2 are continuous random variables with joint pdf $f(y_1, y_2)$, then the marginal pdf's are

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_2 \quad f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_1$$

20.2 Examples

Discrete Example I

A simplified version of an old example is 300 red beads, 200 blue beads mixed in a box. You reach in and randomly select *with replacement* 3 beads. For each red bead, you earn \$1, and for each blue bead, you win \$2. Let R be the number of red beads you select, and W be the winnings you earn. In this case, W is a deterministic function of R, namely

$$W = R + 2(3 - R) = 6 - R$$

but generally, there will be additional randomness in W. The sample space for the experiment is

$$S = \{(r, r, r), (r, r, b), (r, b, r), (b, r, r), (r, b, b), (b, r, b), (b, b, r), (b, b, b)\}$$

The probability of each outcome can be determined by recognizing the independence of each draw:

$$\begin{array}{rr} P((r,r,r)) = \left(\frac{3}{5}\right)^3 = \frac{27}{125} & P((r,r,b)) = \left(\frac{3}{5}\right)^2 \frac{2}{5} = \frac{18}{125} \\ P((r,b,r)) = \left(\frac{3}{5}\right)^2 \frac{2}{5} = \frac{18}{125} & P((b,r,r)) = \left(\frac{3}{5}\right)^2 \frac{2}{5} = \frac{18}{125} \\ P((r,b,b)) = \frac{3}{5}\left(\frac{2}{5}\right)^2 = \frac{12}{125} & P((b,r,b)) = \frac{3}{5}\left(\frac{2}{5}\right)^2 = \frac{12}{125} \\ P((b,b,r)) = \frac{3}{5}\left(\frac{2}{5}\right)^2 = \frac{12}{125} & P((b,b,b)) = \left(\frac{2}{5}\right)^3 = \frac{8}{125} \end{array}$$

Each event maps to one value of R and another value of W (the range of (R, W)). The range of the random variables is

$$(R, W)(S) = \{(3, 3), (2, 4), (1, 5), (0, 6)\}$$

Notice that some events map to the same random bivariate, so the joint pmf is obtained by summing over the mutually exclusive simple events that produce the same random outcome (remember partitions).

$$\begin{array}{ll} p(3,3) = \frac{27}{125} & p(2,4) = \frac{18}{125} + \frac{18}{125} + \frac{18}{125} = \frac{54}{125} \\ p(1,5) = \frac{12}{125} + \frac{12}{125} + \frac{12}{125} = \frac{36}{125} & p(0,6) = \frac{8}{125} \end{array}$$

The marginal pmf for R is obtained directly from the above because there is only one value of w for each value of r, so

$$\begin{array}{ll} p_R(3) = \frac{27}{125} & p_R(2) = \frac{54}{125} \\ p_R(1) = \frac{36}{125} & p_R(0) = \frac{8}{125} \end{array}$$

It should come as no surprise that the above pmf is the same as that of Binomial$(3, 3/5)$,

$$p_R(r) = \binom{3}{r} \left(\frac{3}{5}\right)^r \left(\frac{2}{5}\right)^{3-r}$$

because it is obvious that $R \sim$ Binomial$(3, 3/5)$. Sometimes, when you can directly identify the pmf of the univariate components of the multivariate random variable, you can use this knowledge to verify your calculations of marginal pmfs (or pdfs).

The marginal of W also has the same probabilities, but W does not have a Binomial distribution because its range is $W(S) = \{3, 4, 5, 6\}$.

$$\begin{array}{ll} p_W(3) = \frac{27}{125} & p_W(4) = \frac{54}{125} \\ p_W(5) = \frac{36}{125} & p_W(6) = \frac{8}{125} \end{array}$$

Discrete Example II

The preceding example was rather boring because $W = 6 - R$ was a deterministic function of R. Let's consider a somewhat more interesting example. Suppose there are two contracts to be assigned to three (A, B, or C) firms. Let's suppose that each contract is awarded independently of the other and that all firms are equally likely to win a contract. Let Y_1 be the number of contracts awarded to firm A, and Y_2 the number of contracts awarded to firm B. In this case, while Y_1 certainly informs on Y_2 (for example, if $Y_1 = 2$, then it must be true that $Y_1 = 0$), Y_2 cannot be written as a deterministic function of Y_1.

The sample space of the experiment is

$$S = \{(A,A),(A,B),(B,A),(A,C),(C,A),(B,C),(C,B),(B,B),(C,C)\}$$

and each outcome is equally likely with probability $\frac{1}{9}$ by the problem setup.

The range of the bivariate pair (Y_1, Y_2) is

$$(Y_1, Y_2)(S) = \{(2,0),(1,1),(1,0),(0,1),(0,2),(0,0)\}$$

Clearly, some events map to the same outcome. In particular, the 2nd, 3rd, and 4th outcomes are each mapped to by one simple event. The joint pmf $p(y_1, y_2)$ can be written in table form

		y_1		
		0	1	2
y_2	0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$
	1	$\frac{2}{9}$	$\frac{2}{9}$	0
	2	$\frac{1}{9}$	0	0

The marginal pmf for Y_1 is obtained by summing over columns in this table, i.e. over all possible values of Y_2, so

$$\begin{aligned} p_{Y_1}(0) &= \sum_{y_2} p(0, y_2) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9} \\ p_{Y_1}(1) &= \sum_{y_2} p(1, y_2) = \frac{2}{9} + \frac{2}{9} = \frac{4}{9} \\ p_{Y_1}(2) &= \sum_{y_2} p(2, y_2) = \frac{1}{9} \end{aligned}$$

Again, we can recognize from first principles that Y_1 should have a Binomial$\left(2, \frac{1}{3}\right)$ distribution, which indeed agrees with the calculations above, for example

$$p_{Y_1}(1) = \binom{2}{1}\frac{1}{3} \times \frac{2}{3} = \frac{4}{9}$$

Continuous Example

Recall the specific Dirichlet pdf

$$f(y_1, y_2) = \begin{cases} 360 y_1^2 y_2 (1 - y_1 - y_2) & 0 \le y_1, y_2 \le 1, y_1 + y_2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the marginal pdf $f_{Y_1}(y_1)$.

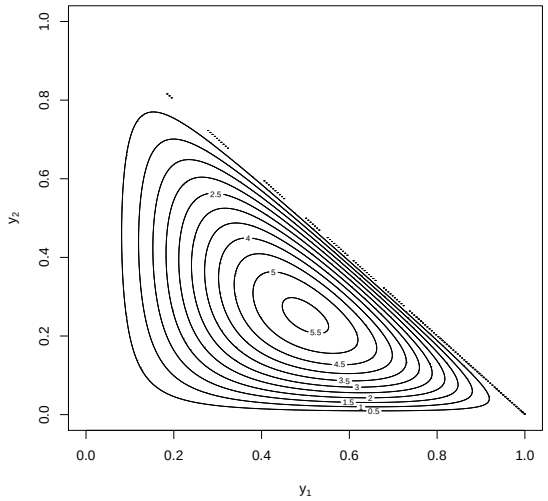


Figure 7: Contour plot of the joint pdf $f(y_1, y_2) = 360y_1^2y_2(1 - y_1 - y_2)$.

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} 360y_1^2y_2(1 - y_1 - y_2)dy_2 \\ &= 360y_1^2 \int_0^{1-y_1} [(1 - y_1)y_2 - y_2^2]dy_2 \\ &= 360y_1^2 \left(\frac{(1 - y_1)y_2^2}{2} - \frac{y_2^3}{3} \right) \Bigg|_0^{1-y_1} \\ &= 360y_1^2 \left(\frac{(1 - y_1)^3}{2} - \frac{(1 - y_1)^3}{3} \right) \\ &= 60y_1^2(1 - y_1)^3 \end{aligned}$$

We recognize this distribution as a Beta$(3, 4)$. Similarly, it can be shown that $f_{Y_2}(y_2)$ is another Beta distribution. Thus, if we slice through the 3-dimensional joint pdf (see Fig. 7), at any value of y_1 (or any value of y_2), we obtain a univariate beta distribution.

21 Conditional Distributions

21.1 Definitions

One of the most important reasons for studying more than one random variable at the same time is to identify relationships between them. Consider the following pairs of random variables and what might be implied by establishing a relationship between them

- Previous mental health evaluation, probability of purchasing a gun.
- Age, probability of being involved in a car accident.
- Skin color, number of loan applications filled out prior to first success.

Definition: *conditional pmf*

Suppose Y_1, Y_2 are discrete random variables with joint pmf $p(y_1, y_2)$ and marginals $p(y_1)$ and $p(y_2)$, then the *conditional probability mass function* is

$$p(y_1 \mid y_2) = P(Y = y_1 \mid Y_2 = y_2) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_2 = y_2)} = \frac{p(y_1, y_2)}{p(y_2)}$$

wherever $p(y_2) > 0$.

Definition: *conditional cdf, pdf*

Suppose Y_1, Y_2 are continuous, then the conditional cdf is

$$F(y_1 \mid y_2) = P(Y_1 \leq y_1 \mid Y_2 = y_2)$$

Furthermore, if they have joint pdf $f(y_1, y_2)$ and marginal pdfs $f(y_1)$ and $f(y_2)$, then the *conditional pdf* is

$$f(y_1 \mid y_2) = \frac{f(y_1, y_2)}{f(y_2)}$$

for all y_2 such that $f(y_2) > 0$.

It would be tempting to state that

$$F(y_1 \mid y_2) = \frac{P(Y_1 \leq y_1, Y_2 = y_2)}{P(Y_2 = y_2)}$$

but it is not valid for continuous random variables because $P(Y_2 = y_2) = 0$. Instead, we interpret the meaning using the usual approximate argument:

$$F(y_1 + \Delta/2 \mid y_2) - F(y_1 - \Delta/2 \mid y_2) = P(y_1 - \Delta/2 < Y_1 \leq y_1 + \Delta/2 \mid y_2) \approx \frac{f(y_1, y_2)\Delta}{f(y_2)}$$

21.2 Examples

1. Returning to our contracts example, we had established the following joint pmf.

		y_1			
		0	1	2	$p(y_2)$
y_2	0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{4}{9}$
	1	$\frac{2}{9}$	$\frac{2}{9}$	0	$\frac{4}{9}$
	2	$\frac{1}{9}$	0	0	$\frac{1}{9}$
	$p(y_1)$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{1}{9}$	

Then, we can derive the y_2 pmf conditional on $y_1 = 0$ as

$$\begin{aligned} p(0 \mid y_1 = 0) &= \frac{p(0,0)}{p_{Y_1}(0)} = \frac{1/9}{4/9} = \frac{1}{4} \\ p(1 \mid y_1 = 0) &= \frac{p(0,1)}{p_{Y_1}(0)} = \frac{2/9}{4/9} = \frac{1}{2} \\ p(2 \mid y_1 = 0) &= \frac{p(0,2)}{p_{Y_1}(0)} = \frac{1/9}{4/9} = \frac{1}{4} \end{aligned}$$

And for $y_1 = 1$, the conditional pmf is

$$\begin{aligned} p(0 \mid y_1 = 1) &= \frac{p(1,0)}{p_{Y_1}(1)} = \frac{2/9}{4/9} = \frac{1}{2} \\ p(1 \mid y_1 = 1) &= \frac{p(1,1)}{p_{Y_1}(1)} = \frac{2/9}{4/9} = \frac{1}{2} \\ p(2 \mid y_1 = 1) &= \frac{p(1,2)}{p_{Y_1}(1)} = \frac{0}{4/9} = 0 \end{aligned}$$

Notice, that as y_1 changes, the conditional pmf changes. Also, notice that for each value of y_1, the conditional pmf describes a proper pmf, i.e. it sums to one over all possible values of y_2.

2. Suppose $N \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Binomial}(N+1, p)$. Find the joint distribution of (N, Y).

   We observe that Y follows a Binomial distribution, *if* the value of random variable N is known. Because the distribution of Y depends on N, we recognize we have been given the conditional distribution $p(y \mid n)$.

   We can construct joint distributions *given* marginal and conditional distributions, using rearrangement

   $$p(y_1, y_2) = p(y_1 \mid y_2)p(y_2)$$

   Thus,

   $$p(n, y) = \binom{n+1}{y} p^y (1-p)^{n+1-y} \frac{e^{-\lambda}\lambda^n}{n!}$$

3. Suppose proportion Y_1 of a gasoline holding tank is filled at the beginning of the week. Then, during the week, customers buy proportion Y_2 of gasoline out of the tank. Clearly, $Y_2 \leq Y_1$ and both exist in the interval $[0, 1]$. We define a joint pdf for both as follows

   $$f(y_1, y_2) = \begin{cases} 3y_1, & 0 \leq y_2 \leq y_1 \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

   (a) Find marginal $f(y_2)$.
   Given y_2, $y_1 \in [y_2, 1]$, so

   $$f(y_2) = \int_{y_2}^{1} 3y_1 dy_1 = \frac{3}{2} y_1^2 \Big|_{y_2}^{1} = \frac{3}{2}\left(1 - y_2^2\right)$$

   We can verify that this is a proper pdf

   $$\int_0^1 f(y_2) dy_2 = \left[\frac{3}{2} y_2 - \frac{1}{2} y_2^3\right]\Big|_0^1 = \frac{3}{2} - \frac{1}{2} = 1.$$

   (b) Find conditional pdf $f(y_1 \mid y_2)$. When is it defined?

   $$f(y_1 \mid y_2) = \frac{f(y_1, y_2)}{f(y_2)} = \frac{3y_1}{\frac{3}{2}\left(1 - y_2^2\right)}$$

   which is defined when $f(y_2) > 0$, i.e. when $y_2 \neq 1$. Of course, $0 \leq y_2 \leq y_1 \leq 1$ still applies.

   (c) Find $P\left(Y_2 > \frac{1}{2} \mid Y_1 = \frac{3}{4}\right)$.
   First, we need $f(y_1) = \int_0^{y_1} 3y_1 dy_2 = 3y_1^2$. Then,

   $$F\left(\frac{1}{2} \middle| \frac{3}{4}\right) = \int_0^{1/2} \frac{f\left(\frac{3}{4}, t_2\right)}{f\left(\frac{3}{4}\right)} dt_2 = \int_0^{1/2} \frac{3 \times \frac{3}{4}}{3 \times \left(\frac{3}{4}\right)^2} dt_2 = \frac{4}{3} \times \frac{1}{2} = \frac{2}{3}$$

22 Independence

22.1 Definitions

Recall that two events A and B are independent if and only if $P(A \cap B) = P(A)P(B)$.

Definition: *independent random variables*

Suppose $Y_1 \sim F(y_1)$ and $Y_2 \sim F(y_2)$ are two random variables with respective cdfs. Furthermore, suppose (Y_1, Y_2) jointly follow joint cdf $F(y_1, y_2)$. Then, Y_1 and Y_2 are said to be *independent* if and only if

$$F(y_1, y_2) = F(y_1)F(y_2)$$

for all (y_1, y_2) in the real plane.

In short, random variables are independent if we can show all events of the form $\{Y_1 \leq y_1\}$ and $\{Y_2 \leq y_2\}$ are independent. Unfortunately, it is not always so easy to show this condition, so there are faster ways to show independence.

22.2 Theorems

Theorem 51. *Discrete random variables Y_1 and Y_2 are independent if and only if*

$$p(y_1, y_2) = p(y_1)p(y_2)$$

for all y_1, y_2. This statement is always true.

Continuous random variables Y_1 and Y_2 are independent if and only if

$$f(y_1, y_2) = f(y_1)f(y_2)$$

for all y_1, y_2. This statement is only true when the required pdfs exist.

Corollary 52. *For discrete random variables, Y_1 and Y_2 are independent if and only if*

$$p(y_1 \mid y_2) = p(y_1)$$

For continuous random variables, Y_1 and Y_2 are independent if and only if

$$f(y_1 \mid y_2) = f(y_1)$$

whenever $f(y_2) > 0$.

And the reverse conditionals also apply.

Proof.

$$p(y_1 \mid y_2) = \frac{p(y_1, y_2)}{p(y_2)} = \frac{p(y_1)p(y_2)}{p(y_2)} = p(y_1)$$

$$f(y_1 \mid y_2) = \frac{f(y_1, y_2)}{f(y_2)} = \frac{f(y_1)f(y_2)}{f(y_2)} = f(y_1)$$

□

Corollary 53. *If Y_1 and Y_2 are continuous with joint pdf $f(y_1, y_2)$ that is defined on finite rectangle $a \leq y_1 \leq b$ and $c \leq y_2 \leq d$, then Then, Y_1, Y_2 are independent if and only if $f(y_1, y_2) = g(y_1)h(y_2)$ where $g(y_1) \geq 0$ and $h(y_2) \geq 0$ but not necessarily pdfs.*

22.3 Examples

These examples demonstrate the two things you should be able to do regarding independent random variables

1. Demonstrate when two (or more) random variables are independent. (For n random variables, the products in the definitions above extend to products of n marginals.)
2. Construct joint pdfs from marginal distributions when you are told the random variables are independent.

Also, at the end is an example of how to compute complex probabilities involving multiple random variables. I think the fact that you now have the ability to compute probabilities like $P(Y_1 > 2Y_2)$, where the left and the right side of the inequality involve random variables, has not yet been well-demonstrated, so study the last example well.

1. Returning to our usual discrete example, we notice that checking independence is equivalent to what we did earlier in the course.

		y_1			
		0	1	2	$p(y_2)$
y_2	0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{4}{9}$
	1	$\frac{2}{9}$	$\frac{2}{9}$	0	$\frac{4}{9}$
	2	$\frac{1}{9}$	0	0	$\frac{1}{9}$
	$p(y_1)$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{1}{9}$	

   It must be true that each entry in the table is given by the product of the two corresponding marginals. Here, we see that

$$p_{y_2}(2) \times p_{y_2}(2) = \frac{1}{9} \times \frac{1}{9} \neq 0$$

   so Y_1 and Y_2 are not independent.

2. Remember the 300 blue beads, 200 red beads, 100 green beads example, yielding 1, 2 or 3 dollars winnings. Let Y_1 be the winnings from the first draw, Y_2 for the second draw. If we draw with replacement, then the probability of each combination is easily computed as a product.

		y_1			
		1	2	3	$p(y_2)$
Y_2	1	$\frac{1}{2} \times \frac{1}{2}$	$\frac{1}{3} \times \frac{1}{2}$	$\frac{1}{6} \times \frac{1}{2}$	$\frac{1}{2}$
	2	$\frac{1}{2} \times \frac{1}{3}$	$\frac{1}{3} \times \frac{1}{3}$	$\frac{1}{6} \times \frac{1}{3}$	$\frac{1}{3}$
	3	$\frac{1}{2} \times \frac{1}{6}$	$\frac{1}{3} \times \frac{1}{6}$	$\frac{1}{6} \times \frac{1}{6}$	$\frac{1}{6}$
	$p(y_1)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	

   The marginals also computed above demonstrate clearly that each entry in the joint pmf is given as a product of corresponding marginals.

   If beads were not drawn with replacement, there would clearly be dependence between the first draw and second draw winnings, as demonstrated below.

		y_1			
		1	2	3	$p(y_2)$
Y_2	1	$\frac{1}{2} \times \frac{299}{599}$	$\frac{1}{3} \times \frac{300}{599}$	$\frac{1}{6} \times \frac{300}{599}$	$\frac{1}{2}$
	2	$\frac{1}{2} \times \frac{200}{599}$	$\frac{1}{3} \times \frac{199}{599}$	$\frac{1}{6} \times \frac{200}{599}$	$\frac{1}{3}$
	3	$\frac{1}{2} \times \frac{100}{599}$	$\frac{1}{3} \times \frac{100}{599}$	$\frac{1}{6} \times \frac{99}{599}$	$\frac{1}{6}$
	$p(y_1)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	

3. Also, returning to example 3, we notice that the product of the marginals

$$f(y_1) = 3y_1^2 \quad f(y_2) = \frac{3}{2}(1 - y_2^2)$$

   does not equal the joint $f(y_1, y_2) = 3y_1$. Therefore, as is intuitively obvious, the proportion of the tank filled is *not* independent of the proportion of the tank used up during the week.

4. Consider the following joint pdf

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{(y_1 - \mu_1)^2}{2\sigma_1^2} - \frac{(y_2 - \mu_2)^2}{2\sigma_2^2}\right]$$

We can integrate this joint pdf twice to find the marginal pdfs. For example,

$$\begin{aligned} f(y_1) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{(y_1-\mu_1)^2}{2\sigma_1^2} - \frac{(y_2-\mu_2)^2}{2\sigma_2^2}\right] dy_2 \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(y_1-\mu_1)^2}{2\sigma_1^2}\right] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{(y_2-\mu_2)^2}{2\sigma_2^2}\right] dy_2 \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(y_1-\mu_1)^2}{2\sigma_1^2}\right] \end{aligned}$$

and we recognize both marginals are Normal distributions. Furthermore, it is trivial to see $f(y_1, y_2) = f(y_1)f(y_2)$. However, we could have recognized this factorization of the joint pdf right away. *If you can factor the joint into a product of two univariate pdfs, then the random variables are independent.*

5. Use the theorems to quickly identify which of these joint pdfs describe independent random variables.

$$f(y_1, y_2) = \begin{cases} 1, & 0 \le y_1, y_2 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Independent with $f(y_1) = f(y_2) = 1$.

$$f(y_1, y_2) = \begin{cases} 3y_1, & 0 \le y_2 \le y_1 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Dependent because range includes a condition $y_2 \le y_1$.

$$f(y_1, y_2) = \begin{cases} e^{-y_1-y_2}, & y_1, y_2 > 0 \\ 0, & \text{otherwise} \end{cases}$$

Independent because e^{-y_1} and e^{-y_2} are exponential pdfs.

$$f(y_1, y_2) = \begin{cases} 4y_1y_2, & 0 \le y_1, y_2 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Independent with $g(y_1) = 4y_1$ and $h(y_2) = y_2$, which are not pdfs.

$$f(y_1, y_2) = \begin{cases} 6(1-y_2), & 0 \le y_1 \le y_2 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Dependent because range includes condition $y_1 \le y_2$.

$$f(y_1, y_2) = \begin{cases} 1, & 0 \le y_1 \le 2, 0 \le y_2 \le 1, 2y_2 \le y_1 \\ 0, & \text{otherwise} \end{cases}$$

Dependent because range includes condition $y_1 \ge 2y_2$.

$$f(y_1, y_2) = \begin{cases} 2, & 0 \le y_1, y_2 \le 1, y_1 + y_2 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Dependent because range includes condition $y_1 + y_2 \le 1$.

$$f(y_1, y_2) = \begin{cases} e^{-y_1}, & 0 \le y_2 \le y_1 \le \infty \\ 0, & \text{otherwise} \end{cases}$$

Dependent because range includes condition $y_2 \le y_1$.

$$f(y_1, y_2) = \begin{cases} y_1 + y_2, & 0 \le y_1, y_2 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Dependent because $y_1 + y_2$ cannot be factored.

$$f(y_1, y_2) = \begin{cases} \frac{y_1}{6} e^{-(y_1+y_2)/2}, y_1, y_2 > 0 \\ 0, \qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$$

Independent because it can be factored into two marginal pdfs, $f(y_2) = \frac{1}{2}e^{-y_2/2}$ (an Exponential) and $f(y_1) = \frac{y_1}{3}e^{-y_1}$ (a Gamma).

6. Suppose Y_1 and Y_2 are independent Exponentially distributed random variables with mean 1. Find $P(Y_1 > Y_2 \mid Y_1 < 2Y_2)$.

   In this case, we are not given a joint pdf, but we can derive one as

$$f(y_1, y_2) = f(y_1) \times f(y_2) = e^{-y_1} \times e^{-y_2} = e^{-y_1 - y_2}$$

   Furthermore, we use the definition of conditional probability

$$P(Y_1 > Y_2 \mid Y_1 < 2Y_2) = \frac{P(Y_1 > Y_2, Y_1 < 2Y_2)}{P(Y_1 < 2Y_2)}$$

   Both the numerator and denominator involve integrating the joint pdf $f(y_1, y_2)$ over specific regions of the (y_1, y_2)-plane. Specifically,

$$P(Y_2 < Y_1 < 2Y_2) = \int_0^\infty \int_{y_1/2}^{y_1} e^{-y_1 - y_2} dy_2 dy_1 = \frac{1}{6}$$

   and

$$P(Y_1 < 2Y_2) = \int_0^\infty \int_{y_1/2}^{\infty} e^{-y_1 - y_2} dy_2 dy_1 = \frac{2}{3}$$

   and the conditional probability becomes

$$P(Y_1 > Y_2 \mid Y_1 < 2Y_2) = \frac{1}{4}$$

23 Expectation

23.1 Definition

Definition: *multivariate expectation*

Suppose $Y_1, \ldots, Y_k$ are discrete with joint pmf $p(y_1, \ldots, y_k)$, then

$$E[g(Y_1, \ldots, Y_k)] = \sum_{y_1} \sum_{y_2} \cdots \sum_{y_k} g(y_1, \ldots, y_k) p(y_1, \ldots, y_k)$$

where the first sum is over all possible values of Y_1, second of Y_2, etc.

Suppose now that the Y_i are continuous with joint pdf $f(y_1, \ldots, y_k)$, then

$$E[g(Y_1, \ldots, Y_k)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, \ldots, y_k) f(y_1, \ldots, y_k) dy_1 \cdots dy_k$$

What is the expectation of $g(Y_1, Y_2) = Y_1$?

I certainly hope it is $E[Y_1]$, so let's check (continuous case):

$$\begin{array}{rcll} E[g(Y_1, Y_2)] & = & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 f(y_1, y_2) dy_1 dy_2 & \text{definition of multivariate expectation} \\ & = & \int_{-\infty}^{\infty} y_1 \int_{-\infty}^{\infty} f(y_1, y_2) dy_2 dy_1 & \text{rearrange order of integration} \\ & = & \int_{-\infty}^{\infty} y_1 f(y_1) dy_1 & \text{definition of marginal pdf} \\ & = & E[Y_1] & \text{definition of univariate expectation} \end{array}$$

Thank goodness!

23.2 Theorems

The multivariate expectation is a linear operator, just as the univariate expectation, so we get some simplifications for linear functions:

Theorem 54.

$$E[a_1 g_1(Y_1,\ldots,Y_k)+\cdots+a_n g_n(Y_1,\ldots,Y_k)] = a_1 E[g_1(Y_1,\ldots,Y_k)]+\cdots+a_n E[g_n(Y_1,\ldots,Y_k)]$$

It is not hard to prove this theorem using properties of integrals, but it is not particularly fun for k dimensions and an arbitrary linear function. The theorem does lead to some immediate simple formulae.

Corollary 55. *For constants a, c, random variables X and Y and functions $g()$ and $h()$, we have*

$$\begin{aligned} E[c] &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} cf(y_1,\ldots,y_k)dy_1\cdots dy_k = c\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(y_1,\ldots,y_k)dy_1\cdots dy_k = c \\ E[ag(X,Y)] &= ag(X,Y) \\ E[g(X)+h(Y)] &= E[g(X)]+E[g(Y)] \\ E[aX+bY] &= aE[X]+bE[Y] \end{aligned}$$

There is a special simplification available *only* for independent random variables.

Theorem 56. *If Y_1 and Y_2 are independent, then*

$$E[g(Y_1)h(Y_2)] = E[g(Y_1)]E[g(Y_2)]$$

Proof.

$$\begin{aligned} E[g(Y_1)h(Y_2)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1)h(y_2)f(y_1,y_2)dy_1dy_2 && \text{definition of multivariate expectation} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1)h(y_2)f(y_1)f(y_2)dy_1dy_2 && \text{independence implies } f(y_1,y_2)=f(y_1)f(y_2) \\ &= \int_{-\infty}^{\infty} h(y_2)f(y_2)\int_{-\infty}^{\infty} g(y_1)f(y_1)dy_1dy_2 && \text{rearrangement} \\ &= \int_{-\infty}^{\infty} h(y_2)f(y_2)E[g(Y_1)]dy_2 && \text{definition of univariate expectation} \\ &= E[g(Y_1)]\int_{-\infty}^{\infty} h(y_2)f(y_2)dy_2 && \text{pull out constant wrt } y_2 \\ &= E[g(Y_1)]E[h(Y_2)] && \text{definition of univariate expectation} \end{aligned}$$

□

23.3 Examples

1. Reconsider the example where proportion Y_1 of a tank is filled at the beginning of the week and proportion Y_2 is purchased during the course of the week. The joint distribution was given by

$$f(y_1,y_2) = \begin{cases} 3y_1, & 0\le y_2\le y_1\le 1 \\ 0, & \text{otherwise} \end{cases}$$

The desired expectations are

$$E[Y_1] = \int_0^1\int_0^{y_1} 3y_1^2 dy_2dy_1 = \int_0^1 3y_1^3dy_1 = \frac{3}{4}$$

and

$$E[Y_2] = \int_0^1\int_0^{y_1} 3y_1y_2 dy_2dy_1 = \int_0^1 \frac{3}{2}y_1^3dy_1 = \frac{3}{8}$$

so

$$E[Y_1-Y_2] = \frac{3}{4}-\frac{3}{8} = \frac{3}{8}$$

Another approach is to first obtain the marginal distributions $f(y_1)$ and $f(y_2)$ and compute the usual univariate expectations.

2. Suppose there are 9 people, 4 married, 3 never married, and 2 divorced. Suppose 3 people are randomly picked. Let Y_1 be the number of married individuals picked and Y_2 the number of never married people picked. Find $E[Y_1Y_2]$.

   The joint pmf is

$$p(y_1, y_2) = \frac{\binom{4}{y_1}\binom{3}{y_2}\binom{2}{3-y_1-y_2}}{\binom{9}{3}}$$

   for $0 \le y_1, y_2 \le 3$ and $1 \le y_1 + y_2 \le 3$. (There are other ways to come up with this pmf.) We can arrange all the possibilities into a table.

$\times \frac{1}{\binom{9}{3}}$		Y_1			
		0	1	2	3
Y_2	0	0	4	12	4
	1	3	24	18	0
	2	6	12	0	0
	3	1	0	0	0

   The expectation is obtained by multiplying each entry by y_1y_2 and summing over the rows and columns.

$$E[Y_1Y_2] = \frac{1}{\binom{9}{3}}(24 + 2 \times 18 + 2) = 1$$

   Another expectation (did in class):

$$E[Y_1] = \frac{1}{\binom{9}{3}}(4 + 24 + 12 + 2 \times 12 + 2 \times 18 + 3 \times 4) = \frac{112}{84} = \frac{4}{3}$$

3. Consider the following joint pdf and use it to compute $E[Y_1Y_2]$.

$$f(y_1, y_2) = \begin{cases} 6y_1^2y_2, & 0 \le y_1, y_2 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

   Because $f(y_1, y_2) = 6y_1^2 \times y_2$ can be factored and the range is a finite rectangle, we conclude Y_1 and Y_2 are independent. Therefore, $E[Y_1Y_2] = E[Y_1]E[Y_2]$. This simplification is only helpful if we also have the marginal distributions, so we can compute the univariate, marginal expectations, but the marginals can be identified. Both are Beta distributions: $Y_1 \sim \text{Beta}(3, 1)$ and $Y_2 \sim \text{Beta}(2, 1)$. These have easily obtained expectations $\frac{\alpha}{\alpha+\beta}$, in this case $E[Y_1] = \frac{3}{4}$ and $E[Y_2] = \frac{2}{3}$, so $E[Y_1Y_2] = \frac{1}{2}$.

4. Suppose Y_1 and Y_2 are independent and $Y_1 \sim \chi^2_{\nu_1}$ and $Y_2 \sim \chi^2_{\nu_2}$ have chi-squared distributions with ν_1 or ν_2 degrees of freedom. Recall that $E[Y] = \nu$ and $V(Y) = 2\nu$ if $Y \sim \chi^2_\nu$. Find $E[Y_1 + Y_2]$ and $V(Y_1 + Y_2)$.

   Very easily, we use the linearity of expectation to conclude $E[Y_1 + Y_2] = E[Y_1] + E[Y_2] = \nu_1 + \nu_2$.

   For the variance, we start with the definition

$$\begin{aligned}
V(Y_1 + Y_2) &= E[(Y_1 + Y_2 - E[Y_1 + Y_2])^2] && \text{definition of variance} \\
&= E[(Y_1 + Y_2 - \nu_1 - \nu_2)^2] && \text{using the first result} \\
&= E[(Y_1 + Y_2)^2 - 2(\nu_1 + \nu_2)(Y_1 + Y_2) + (\nu_1 + \nu_2)^2] && \text{algebraic expansion} \\
&= E[(Y_1 + Y_2)^2] - 2(\nu_1 + \nu_2)E[Y_1 + Y_2] + (\nu_1 + \nu_2)^2 && \text{linearity of expectation} \\
&= E[(Y_1 + Y_2)^2] - (\nu_1 + \nu_2)^2 && \text{use first result again} \\
&= E[Y_1^2 + 2Y_1Y_2 + Y_2^2] - (\nu_1 + \nu_2)^2 && \text{expansion} \\
&= E[Y_1^2] + 2E[Y_1Y_2] + E[Y_2^2] - (\nu_1 + \nu_2)^2 && \text{linearity of expectation} \\
&= (2\nu_1 + \nu_1^2) + 2E[Y_1Y_2] + (2\nu_2 + \nu_2^2) - (\nu_1 + \nu_2)^2 && V(Y_i) = E[Y_i^2] - (E[Y_i])^2 \\
&= (2\nu_1 + \nu_1^2) + 2E[Y_1]E[Y_2] + (2\nu_2 + \nu_2^2) - (\nu_1 + \nu_2)^2 && \text{independence} \\
&= (2\nu_1 + \nu_1^2) + 2\nu_1\nu_2 + (2\nu_2 + \nu_2^2) - (\nu_1 + \nu_2)^2 && \text{replace } E[Y_i] \\
&= 2\nu_1 + 2\nu_2 && \text{simplify}
\end{aligned}$$

Please note, you will see a *much* easier way to compute this variance of a sum of random variables very soon. The point of illustrating this approach is to get you familiar with working with multivariate expectations, of which variance is a particular example.

5. Non-independent exponentials. Consider the following joint pdf

$$f(y_1, y_2) = \begin{cases} [1 - \alpha(1 - 2e^{-y1})(1 - 2e^{-y2})]e^{-y_1 - y_2}, & y_1, y_2 \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Find $E[Y_1], E[Y_2], V(Y_1), V(Y_2), E[Y_1 - Y_2], E[Y_1 Y_2], V(Y_1 - Y_2)$.

One can derive the marginal distributions to show that both $Y_1, Y_2 \sim \text{Exponential}(1)$ have exponential distributions. They are *not*, however, independent because clearly $f(y_1) = e^{-y_1}$ and $f(y_2) = e^{-y_2}$ and $f(y_1, y_2) \neq f(y_1)f(y_2)$.

It is trivial to show $E[Y_1] = E[Y_2] = V(Y_1) = V(Y_2) = 1$ and therefore $E[Y_1 - Y_2] = 0$.

$$\begin{array}{rcll} E[Y_1Y_2] & = & \int_0^\infty \int_0^\infty y_1 y_2 [1 - \alpha(1 - 2e^{-y_1})(1 - 2e^{-y_2})] e^{-y_1 - y_2} dy_1 dy_2 & \text{multi. expect.} \\ & = & (1 - \alpha) \int_0^\infty \int_0^\infty y_1 y_2 e^{-y_1 - y_2} + 2\alpha \int_0^\infty \int_0^\infty y_1 y_2 e^{-2y_1 - y_2} dy_1 dy_2 & \\ & & + 2\alpha \int_0^\infty \int_0^\infty y_1 y_2 e^{-2y_1 - y_2} dy_1 dy_2 - 4\alpha \int_0^\infty \int_0^\infty y_1 y_2 e^{-2y_1 - 2y_2} dy_1 dy_2 & \\ & = & (1 - \alpha)\Gamma(2) + \alpha\Gamma(2) - \frac{\alpha\Gamma(2)}{16} & \text{integrals are Gamma functions} \\ & = & 1 - \frac{\alpha}{4} & \Gamma(2) = 1 \end{array}$$

By the same argument as before, and using $E[Y_1Y_2]$ from above (can no longer use $E[Y_1Y_2] = E[Y_1]E[Y_2]$ because of non-independence, we have

$$\begin{aligned} V(Y_1 - Y_2) &= E[Y_1^2] - 2E[Y_1Y_2] + E[Y_2^2] - E[Y_1 - Y_2] \\ &= (1 + 1^2) - 2\left(1 - \frac{\alpha}{4}\right) + (1 + 1^2) - 0 \\ &= 2 + \frac{\alpha}{2} \end{aligned}$$

24 Covariance

24.1 Definitions

We have now dealt with dependent random variables. We are often interested in characterizing the nature of the dependence between them. We had a fairly lengthy discussion in class about what that dependence may look like. We justified that one summary of a *linear* dependence is the covariance, as defined below. Recognizing and visualizing dependences between pairs of random variables is a very important concept. You may wish to look at some of the pictures here.

Definition: *covariance*

If Y_1 and Y_2 are two random variables, then the covariance is defined as

$$\text{Cov}(Y_1, Y_2) = E[(Y_1 - E[Y_1])(Y_2 - E[Y_2])]$$

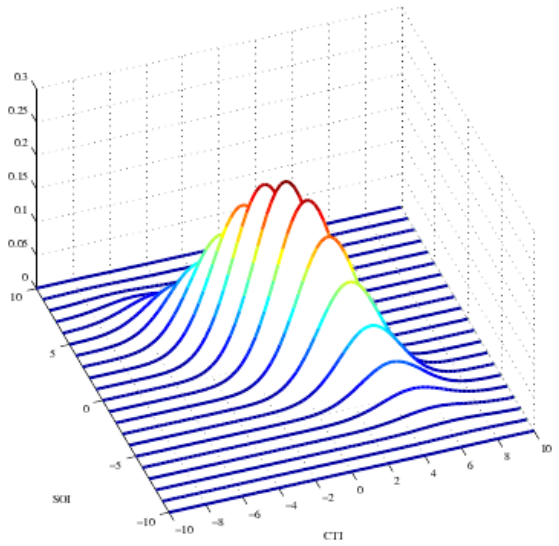
Properties:

1. Let $E[Y_1] = \mu_1$ and $E[Y_2] = \mu_2$. If Y_1 and Y_2 fall very close to a line with positive slope, then for the usual random point (Y_1, Y_2), the product $(Y_1 - \mu_1)(Y_2 - \mu_2)$ will tend to be positive because either both are positive or both are negative, but very rarely will one be negative and the other positive. Thus, covariance will be positive.

2. If Y_1 and Y_2 fall very close to a line with negative slope, then the product will tend to be negative and covariance will be negative.

3. If Y_1 and Y_2 are independent, then the distribution of y_2, for example, is not changed by the value of y_1, that is $f(y_2 \mid y_1) = \frac{f(y_1,y_2)}{f(y_1)} = \frac{f(y_1)f(y_2)}{f(y_1)} = f(y_2)$. This implies that there is no discernible relationship between Y_1 and Y_2 and it is difficult to predict the sign of $(Y_1 - \mu_1)(Y_2 - \mu_2)$. In fact, sometimes it will be positive, sometimes negative, and *on average* it will be about 0.

If you are a visual person, you really need to draw pictures of relationships between random variables. For example, the following plot shows a *positive linear relationship* between Y_1 and Y_2. Notice, that unlike mathematical functions for pairs (x, y), say function $h(x)$ that maps each x to the corresponding $y = h(x)$, relationships among random variables are only approximate. If $g(y) = mx + b$ is a line with positive slope that characterizes the relationship between Y_1 and Y_2, then Y_1 maps to $Y_2 = g(Y_1) + \epsilon$ where ϵ is some random term that accounts for the sloppiness (randomness) of the relationship.



There is a problem with covariance as a measure of the linearity of the relationship between two random variables. If I suddenly change the scale of Y_2, say let $Y_2' = 10Y_2$, then the covariance $\text{Cov}(Y_1, Y_2)$ will be bigger by a factor of 10 (see the definition). Well, the covariance may be bigger, but the strength of the relationship between Y_1 and Y_2 is identical to the strength of the relationship between Y_1 and Y_2'. What gives?

Simple, covariance is not the best measure of linearity in a random relationship. Instead, we define Definition: *correlation coefficient*

If Y_1 and Y_2 are two random variables, then

$$\rho = \frac{\text{Cov}(Y_1, Y_2)}{\sigma_1 \sigma_2}$$

where $\sigma_1 = \sqrt{V(Y_1)}$ and similar for σ_2.

24.2 Theorems

Theorem 57.

$$\rho \in [-1, 1]$$

and ρ is invariant to scaling.

Now, it is possible to characterize the strength of a linear relationship. If (Y_1, Y_2) have $\rho_Y = 0.87$ and (X_1, X_2) have $\rho_X = 0.63$, then the relationship between Y_1 and Y_2 is more linear than the X's.

Theorem 58.

$$\textit{Cov}(Y_1, Y_2) = E[Y_1 Y_2] - E[Y_1]E[Y_2]$$

The above theorem makes calculation of covariance easier, much like the previous result $V(Y) = E[Y^2] - (E[Y])^2$ for variances.

Theorem 59. *If Y_1 and Y_2 are independent, then Cov$(Y_1, Y_2) = 0$.*

Proof.

$$\text{Cov}(Y_1, Y_2) = E[Y_1 Y_2] - E[Y_1]E[Y_2] = E[Y_1]E[Y_2] - E[Y_1]E[Y_2] = 0$$

□

Please note, this theorem is a forward implication *only*, so it does not imply that if $\text{Cov}(Y_1, Y_2) = 0$ that then Y_1, Y_2 are independent.

The definition of covariance allows a result that will help you deal with variance computations for linear functions of multiple random variables.

Theorem 60.

$$V(a_1 Y_1 + \cdots + a_n Y_n) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2\sum\sum_{1 \le i < j \le n} a_i a_j \textit{Cov}(Y_i, Y_j)$$

The proof involves a lot of unfriendly algebra, but as we've seen in previous examples, expands the squared difference and uses the fact that the multivariate expectation is a linear function. The result is very helpful. Some concrete examples of this formula are:

$$\begin{aligned} V(Y_1 + Y_2) &= V(Y_1) + V(Y_2) + 2\text{Cov}(Y_1, Y_2) \\ V(Y_1 - Y_2) &= V(Y_1) + V(Y_2) - 2\text{Cov}(Y_1, Y_2) \\ V(2Y_1 + 3Y_2 - Y_3) &= 4V(Y_1) + 9V(Y_2) + V(Y_3) + 12\text{Cov}(Y_1, Y_2) - 4\text{Cov}(Y_1, Y_3) - 6\text{Cov}(Y_2, Y_3) \end{aligned}$$

24.3 Examples

1. Let's revisit the dependent exponentials example with

$$f(y_1, y_2) = \begin{cases} [1 - \alpha(1 - 2e^{-y1})(1 - 2e^{-y2})]e^{-y_1 - y_2}, & y_1, y_2 \ge 0 \\ 0, & \text{otherwise} \end{cases}$$

   We determined that $E[Y_1 Y_2] = 1 - \alpha/4$, so we can immediately compute

$$\text{Cov}(Y_1, Y_2) = E[Y_1 Y_2] - E[Y_1]E[Y_2] = 1 - \alpha/4 - 1 \times 1 = -\frac{\alpha}{4}$$

   Notice that if $\alpha = 0$, then $\text{Cov}(Y_1, Y_2) = 0$. Because the theorem about independence is a forward implication only, this does *not* necessarily mean that Y_1 and Y_2 are independent. However, in this case, we notice that $\alpha = 0$ leads to joint pdf

$$f(y_1, y_2) = \begin{cases} e^{-y_1 - y_2}, & y_1, y_2 \ge 0 \\ 0, & \text{otherwise} \end{cases}$$

   which we have previously established does imply independence of Y_1 and Y_2.

2. $\text{Cov}(c, Y) = E[(c - c)(Y - E[Y])] = E[0 \times (Y - E[Y])] = 0$. So the covariance of a constant and any random variable is 0.

3. Consider the joint uniform random variables on a specific triangle

$$f(y_1, y_2) = c \begin{cases} 1, & -1 \le y_1 \le 0, 0 \le y_2 \le y_1 + 1 \\ 1, & 0 \le y_1 \le 1, 0 \le y_2 \le 1 - y_1 \\ 0, & \text{otherwise} \end{cases}$$

Because the triangle in question has unit area, it is clear that $c = 1$. What is $\text{Cov}(Y_1, Y_2)$?

We will make use of the expression $E[Y_1Y_2] - E[Y_1]E[Y_2]$. First, the marginal distribution of Y_1 is

$$f(y_1) = \begin{cases} \int_0^{y_1+1} dy_2 = y_1 + 1, & -1 \le y_1 \le 0 \\ \int_0^{1-y_1} dy_2 = 1 - y_1, & 0 \le y_1 \le 1 \end{cases}$$

so

$$E[Y_1] = \int_{-1}^{0} (y_1^2 + y_1) dy_1 = \int_0^1 (y_1 - y_1^2) dy_1 = -\left(-\frac{1}{3} + \frac{1}{2}\right) + \frac{1}{2} - \frac{1}{3} = 0$$

There is no need to compute $E[Y_2]$ because it will be eliminated by this 0 in the final expression. We do need

$$E[Y_1Y_2] = \int_0^1 \int_{y_2-1}^{1-y_2} y_1 y_2 dy_1 dy_2 = \int_0^1 y_2 \left(\frac{(1-y_2)^2}{2} - \frac{(y_2-1)^2}{2}\right) dy_2 = 0$$

Because of the symmetry of the region and the function y_1y_2 around $y_1 = 0$, this integral evaluates to 0, thus

$$\text{Cov}(Y_1, Y_2) = 0.$$

However, this time Y_1 and Y_2 are *not* independent. We can see this even without computing $f(y_2)$. It is clear that

$$f(y_1, y_2) = \begin{cases} 1 \\ 0 \end{cases} \neq f(y_1)f(y_2) = \begin{cases} (y_1+1)f(y_2), & -1 \le y_1 \le 0 \\ (1-y_1)f(y_2), & 0 \le y_1 \le 1 \end{cases}$$

It would be necessary for $f(y_2) = \frac{1}{1+y_1}$ in some part of its domain, but that is not possible, as $f(y_2)$ is not a function of y_1.

4. Let's apply the final theorem about variance of linear sums of random variables to previous examples.

For the case when Y_1 and Y_2 were independent chi-square distributions, we see

$$V(Y_1 + Y_2) = V(Y_1) + V(Y_2) = 2\nu_1 + 2\nu_2$$

The covariance term is 0 because of independence, and we get the same result much easier.

For the case where Y_1 and Y_2 were dependent exponentials, we see

$$V(Y_1 - Y_2) = V(Y_1) + V(Y_2) - 2\text{Cov}(Y_1, Y_2) = 1 + 1 - 2\frac{-\alpha}{4} = 2 + \frac{\alpha}{2}$$

Sigh, so much easier!