Course notes from MITx 14.310x Data Analysis for Social Scientists (EdX)

James Solomon-Rounce

2018-10-17

Contents

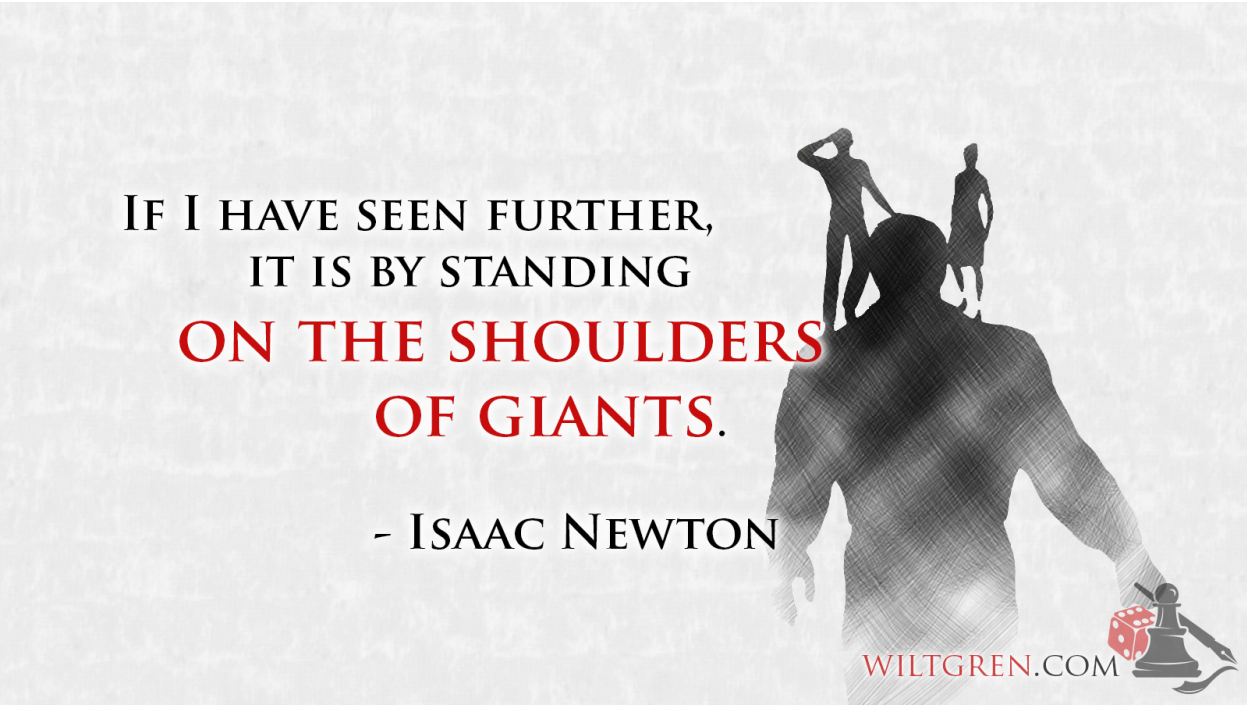
Preface

The following notes were taken by me for educational, non-commercial, purposes. If you find the information useful, buy the material/take the course.

Thank you to the original content providers. Additional ramblings are my own.

Core Resources

- Course Schedule
- Grading and Homework Policy
- Honor Code and Collaboration Guide
- Notes - OLS Derivation
- Notes - Matrix Notation
- R Studio Cheatsheets
- R for Data Science Book



Chapter 1

Module 1: Introduction to the Course

Module Sections:

- Welcome to the Course
- Introduction to R
- Introductory Lecture - Data is Beautiful, Insightful, Powerful, Deceitful
- Finger Exercises
- Module 1: Homework

Module Content:

- Module 1 Slides
- Homework 1 Background Paper - The Persistent Effects Of Peru's Mining Mita
- R Instructions

- Intro to R Zip File

- Citations Data for Homework 1

1.1 Introduction to R

First we setup the environment and install the course files

```
library(swirl)
install_course_zip("./files/M1/14_310x_Intro_to_R_.zip",multi=FALSE)
swirl()
```

IF z is a three number vector e.g.

```
z <- c(1.1, 9, 3.14)
```

If we take the square root of z - 1 and assign it to a new variable called my_sqrt e.g.

```
my_sqrt <- sqrt(z - 1)
```

The result is a vector of length three e.g.

```
my_sqrt
```

```
## [1] 0.3162278 2.8284271 1.4628739
```

Next, if we create a new variable called my_div that gets the value of z divided by my_sqrt.

```
my_div <- z / my_sqrt
```

The first element of my_div is equal to the first element of z divided by the first element of my_sqrt, and so on…

```
my_div
```

```
## [1] 3.478505 3.181981 2.146460
```

When given two vectors of the same length, R simply performs the specified arithmetic operation (+, -, *, etc.) element-by-element. If the vectors are of different lengths, R 'recycles' the shorter vector until it is the same length as the longer vector.

If the length of the shorter vector does not divide evenly into the length of the longer vector, R will still apply the 'recycling' method, but will throw a warning.

```
c(1, 2, 3, 4) + c(0, 10, 100)
```

```
## Warning in c(1, 2, 3, 4) + c(0, 10, 100): longer object length is not a
## multiple of shorter object length
```

```
## [1]   1  12 103   4
```

1.1.1 Module 1 Homework

This is a sample of some of the homework answers. Some questions were observational or required interpretation of maps for example, as such these are not inluded here.

```
library(tidyverse)
```

```
## -- Attaching packages ------------------ tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts --------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
papers <- as_tibble(read_csv("./files/M1/CitesforSara.csv"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   journal = col_character(),
##   title = col_character(),
##   au1 = col_character(),
##   au2 = col_character(),
##   au3 = col_character(),
##   past5 = col_double(),
##   aflpn90 = col_double(),
##   aulpn90 = col_double(),
##   aulpn80 = col_double(),
##   aulpn70 = col_double(),
##   lcites = col_double()
```

```
## )

## See spec(...) for full column specifications.
```

Q. 19 Let's take a look at some of the most popular papers. Using the filter() method, how many records exist where there are greater than or equal to 100 citations?

```
#First lets look at our data
head(papers)
```

```
## # A tibble: 6 x 22
##   journal  year cites title au1   au2   au3   female1 female2 female3  page
##   <chr>   <int> <int> <chr> <chr> <chr> <chr>   <int>   <int>   <int> <int>
## 1 Americ~  1993    31 Jeux~ Kanb~ Keen~ <NA>        0       0      NA    16
## 2 Americ~  1993     4 Chan~ Jame~ <NA>  <NA>        0      NA      NA    22
## 3 Americ~  1993    28 Fact~ Bert~ <NA>  <NA>        0      NA      NA    15
## 4 Americ~  1993    10 Stra~ Garf~ Oh,-~ <NA>        1       0      NA    19
## 5 Americ~  1993     5 Will~ Coat~ Lour~ <NA>        0       0      NA    21
## 6 Americ~  1993    21 Merg~ Kim,~ Sing~ <NA>        0       0      NA    21
## # ... with 11 more variables: order <int>, nauthor <int>, past5 <dbl>,
## #   aflpn90 <dbl>, spage <int>, field <int>, subfld <int>, aulpn90 <dbl>,
## #   aulpn80 <dbl>, aulpn70 <dbl>, lcites <dbl>
```

```
arrange(papers,desc(cites), title)
```

```
## # A tibble: 4,182 x 22
##    journal  year cites title au1   au2   au3   female1 female2 female3
##    <chr>   <int> <int> <chr> <chr> <chr> <chr>   <int>   <int>   <int>
##  1 Econom~  1980  2251 A He~ Whit~ <NA>  <NA>        0      NA      NA
##  2 Econom~  1979  2227 Pros~ Kahn~ Tver~ <NA>        0       0      NA
##  3 Econom~  1987  2164 Co-i~ Engl~ Gran~ <NA>        0       0      NA
##  4 Econom~  1979  1602 Samp~ Heck~ <NA>  <NA>        0      NA      NA
##  5 Econom~  1978  1217 Spec~ Haus~ <NA>  <NA>        0      NA      NA
##  6 Econom~  1982  1077 Auto~ Engl~ <NA>  <NA>        0      NA      NA
##  7 Econom~  1981  1031 Like~ Dick~ Full~ <NA>        0       0      NA
##  8 Econom~  1982   983 Larg~ Hans~ <NA>  <NA>        0      NA      NA
##  9 Econom~  1980   864 Macr~ Sims~ <NA>  <NA>        0      NA      NA
## 10 Econom~  1982   563 Time~ Kydl~ Pres~ <NA>        0       0      NA
## # ... with 4,172 more rows, and 12 more variables: page <int>,
## #   order <int>, nauthor <int>, past5 <dbl>, aflpn90 <dbl>, spage <int>,
## #   field <int>, subfld <int>, aulpn90 <dbl>, aulpn80 <dbl>,
## #   aulpn70 <dbl>, lcites <dbl>
```

```
papers %>%
  filter(cites >= 100)
```

```
## # A tibble: 205 x 22
##    journal  year cites title au1   au2   au3   female1 female2 female3
##    <chr>   <int> <int> <chr> <chr> <chr> <chr>   <int>   <int>   <int>
##  1 Americ~  1994   117 Is I~ Pers~ Tabe~ <NA>        0       0      NA
##  2 Econom~  1971   149 Furt~ Nerl~ <NA>  <NA>        0      NA      NA
##  3 Econom~  1971   170 The ~ Madd~ <NA>  <NA>       NA      NA      NA
##  4 Econom~  1971   155 Inve~ Luca~ Pres~ <NA>        0       0      NA
##  5 Econom~  1971   139 Some~ Crag~ <NA>  <NA>        0      NA      NA
##  6 Econom~  1971   108 Iden~ Roth~ <NA>  <NA>        0      NA      NA
##  7 Econom~  1972   164 Meth~ Fair~ Jaff~ <NA>        0       0      NA
##  8 Econom~  1972   150 Exis~ Radn~ <NA>  <NA>        0      NA      NA
```

```
##  9 Econom~  1973   361 Mani~ Gibb~ <NA>  <NA>        0      NA      NA
## 10 Econom~  1973   107 On a~ Kram~ <NA>  <NA>        0      NA      NA
## # ... with 195 more rows, and 12 more variables: page <int>, order <int>,
## #   nauthor <int>, past5 <dbl>, aflpn90 <dbl>, spage <int>, field <int>,
## #   subfld <int>, aulpn90 <dbl>, aulpn80 <dbl>, aulpn70 <dbl>,
## #   lcites <dbl>
```

Q.20 Use the group_by() function to group papers by journal. How many total citations exist for the journal "Econometrica"?

```
papers %>%
  group_by(journal) %>%
  summarise(sum(cites))
```

```
## # A tibble: 7 x 2
##   journal                           `sum(cites)`
##   <chr>                                    <int>
## 1 American-Economic-Review                  3738
## 2 Econometrica                             75789
## 3 Journal-of-Political-Economy              3398
## 4 Quarterly-Journal-of-Economics            8844
## 5 Review-of-Economic-Studies               21937
## 6 Review-of-Economics-and-Statistics        8473
## 7 <NA>                                        14
```

```
#or

summarize(group_by
          (papers, journal),
          SumOfCites = sum(cites))
```

```
## # A tibble: 7 x 2
##   journal                           SumOfCites
##   <chr>                                  <int>
## 1 American-Economic-Review                3738
## 2 Econometrica                           75789
## 3 Journal-of-Political-Economy            3398
## 4 Quarterly-Journal-of-Economics          8844
## 5 Review-of-Economic-Studies             21937
## 6 Review-of-Economics-and-Statistics      8473
## 7 <NA>                                      14
```

Q.21 How many distinct primary authors (au1) exist in this dataset?

```
papers %>%
  summarise(n_distinct(au1))
```

```
## # A tibble: 1 x 1
##   `n_distinct(au1)`
##               <int>
## 1              2332
```

```
#or

n_distinct(papers$au1)
```

```
## [1] 2332
```

Chapter 2

Module 2: Fundamentals of Probability, Random Variables, Joint Distributions + Collecting Data

---

Module Sections:

- Fundamentals of Probability
- Random Variables, Distributions, and Joint Distributions
- Gathering and Collecting Data
- Module 2: Homework

Module Content:

- Section 2 Slides - Fundamentals of Probability
- Section 3 Slides - Random Variables, Distributions and Joint Distributions
- Section 4 Slides - Gathering and Collecting Data

2.1 Fundamentals of Probability

2.1.1 Set Theory

- A *sample space* is collection of all possible outcomes
- An *event* is any collection of outcomes - could be one, all or none
- If the outcome is a member of an event, the event is said to have *occured*
- Event B is said to be *contained* by A, if all outcomes in B also are in A
- This is the basis of set theory and used widely in probability, although there are some differences between set and probability theory
- If there is no symbol, then this usually means intersection AB in probability - in set theory we would write an inverted U e.g. $A \cap B$
- A and B are mutually exclusive (disjoint in set theory) if they have no outcomes in common
- A and B are exhaustive (complimentary in set theory) if their union is S (the entire sample space)

- A and B are both mutually exclusive and exhaustive, their union is equal to the sample space but they have no events in common - they are a partition of the sample space

2.1.2 Defining Probability

We assign every event a number P(A) which is the prob. the event will occur

1 We require that the probability is greater than one for all events in the sample space - P(A) >= 0 for all A c S 2 The entire sample space must be equal to one - P(S) = 1 3 For any sequence of disjoint sets, the prob. of the union of that sequence is equal to the sum of the probabilities of those events - A_1, A_2, … , is $P(V_i) = \sum_i P(A_i)$

So we have a sample space, and if it satisfies these three properties, then we call it a probability. Sometimes this is referred to as a probability function or a probability distribution, but there is no standard terminology for all probability theory. Set theory helps to prove aspects of probability mathematically, for the purposes on this course, we just need to know what some useful facts are.

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

The probability of A compliment, which is the event that contains all of the outcomes that are not in the event A, the probability of A compliment is just equal to 1 minus the probability of A. This is useful if the probability of A comliment ($P(A^c)$) is difficult to compute, where as the probability of A might be very easy to compute.

- $P(\emptyset)$ =

The probability of the empty set is zero.

- If A $_c$ B then P(A) <= P(B) =

If A is contained in B then the probability of A is less than or equal to the probability of B

- For all A, 0 <= P(A) <= 1 =

For any events, the probability of that event is between 0 and 1.

- P(AUB) = P(A) + P(B) - P(AB) =

Probability of A union B is just equal to the sum of the probabilities of those two events minus the probability of the their intersection.

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

The probability of A times B complement is equal to the probability of A minus the probability of the intersection.

2.1.3 An example

Suppose you have a finite sample space. Let the function n(.) give the number of elements in a set.

Then define P(A) = n(A)/n(S). This is called a simple sample space, and it is a probability - we count the number of outcomes and divide by the number of possible outcomes in the sample space.

We can check that it satisfies the three axioms to ensure it is a probability:

1. P(A) will always be non-negative because it's a count
2. P(S) will equal 1, by definition
3. P(AUB) = n(AUB)/n(S) = n(A)/n(S) + n(B)/n(S) = P(A) +P(B).

If you can put your experiment in to this sample space where each outcome is equally likely, we just need to count to calculate probabilities of events. So for example, if you want to know how likely it is you will roll a specific number, say 6, on two dice, we calculate all the different ways that six occurs then divide this by all possible outcomes (sample space) - = 5 / 36 = or 13.9%

2.1.4 Another example

If the state of Massachusetts issues 6-character license plates, using one of 26 letters and 10 digits randomly for each character, what is the probability that I will receive an all digit license plate?

n(S) = 36 (26 + 10) possibilities for each of 6 characters = 36^6 = 2.176b n(A) = 10 possibilities (for digits only) for each of 6 characters = 10^6 = 1m so P(A) = .0005

This is *sampling with replacement*

What if Massachusetts does not reuse a letter or digit?

Now, in the sample space, there are 36 possibilities (26 + 10) for the 1st character, 35 left for the 2nd, and so on.

n(S) = 36x35x34x33x32x31 = 36!/30! = 1.402b

Similarly, in the event, there are 10 possibilities for the 1st character, 9 left for the 2nd, and so on.

n(A) = 10x9x8x7x6x5 = 10!/4! = 151k

so P(A) = 1.402b / 151k = .0001

This is *sampling without replacement*

2.1.5 Ordered and Unordered Arrangements

In the examples so far, we have used a series of counting rules - combinatorics i.e. combinations of objects belonging to a finite set in accordance with certain constraints.

1. If an experiment has two parts, first one having m possibilities and, regardless of the outcome in the first part, the second one having n possibilities, then the experiment has m * n possible outcomes - this is what we do intuitively
2. Any ordered arrangement of objects is called a *permutation*. The number of different permutations of N objects is N! The number of different permutations of n objects taken from N objects is N!/(N-n)! This is the case in the license plate example previously given
3. Any unordered arrangement of objects is called a *combination*. The number of different combinations of n objects taken from N objects is N!/{(N-n)!n!}. We typically denote this $\binom{N}{n}$ - N (big objects) choose n (combinations). This is where the order of objects doesn't matter i.e. different orderings don't matter - we take out the ordering

So if we had 9 people who each wanted to shake hands, if order doesn't matter then it is a combination and we take 9 and choose 2 so it becomes:

9! / {(9-2)! * 2!} = 9! / {7! * 2!} = 362,880 / {5,040 * 2} = 362,880 / 10,080 = 36 combinations

Note, if order did matter and we used the permutations formula the total would be twice as many

2.1.6 Office Arrangements and Pizza Toppings

Q: If there are six vegetarian pizza toppings and five non-veg, if I randomly choose two from a hat containing all items, what is the probability that I end up with a pizza that has one veg and one non-veg topping?

A:

First we need to count the number of possabilities in the sample space e.g. {(V1, V2), (V1, V3), (V1, V4), (V1, N1) ...} n(S) = $\binom{11}{2}$ = 55 - All outcomes are equally likely

Now we need to define our outcome n(A) = there are A = {(V1, N1), (V1, N2), (V2, N1) ... } n(A) = 6 * 5 = 30

So the probability is N(A) / n(S) = 30 / 55 = 0.55

In general, I could have chose n toppings and asked what is the probability that my pizza had n1 vegetarian toppings and n2 non-vegetarian toppings. There would, then, be $\binom{6}{n_1}$ possibilities for the veg toppings and $\binom{5}{n_2}$for the non-veg toppings. In other words,

$$P(n_1 veg, n_2 non-veg) = \binom{6}{n_1}\binom{5}{n_2}\binom{11}{n}$$

This is the basis of the hypergeometric distribution.

2.1.7 Independence and Basketball Example

We call probabilistic events stochastic events. One of the most fundamental relationships between stochastic events is independence.

- Events A and B are independent if P(AB) = P(A) P(B)

That is to say, events A and B are independent if the probability of their intersection is equal to the product of their probabilities.

- independent events is that knowing one event occurred doesn't give you any information about whether the other occurred.

This is best represented with an example. If you toss one die, once. Consider the event, A, that you roll a number less than 5, and the event, B, that you roll an even number. Are these events independent?

You might consider how could they be, as they rely on the same roll of a die?

If we use the previous example for independence, we check:

1. Probability of event A is P(A) = 2/3
2. Probability of event B is P(B) = 1/2
3. Probability of their intersection is P(AB) = 1/3 which is the same as P(A) P(B)

So yes, it does satisfy the definition of independence. AB is rolling an even number less than 5 (e.g. 2 or 4) and P(A)P(B) = P(AB)

So knowing one event occurred doesn't give you any information about whether an other occured

In another example, if we had a bag of ten poker chips numbered 1 to 10, with 3 different colours - Red(1,2,3,4,5), Blue(6,7) or Green(8,9,10)

If choosing a poker chip, A that it is blue, and B that it is even, independent?

1. Probability of event A is P(A) = 2/10 (.2)
2. Probability of event B is P(B) = 5/10 (.5)
3. Probability of their intersection is P(AB) = 1/10 (or .1) which is the same as P(A) P(B)

So yes they are independent, knowing one (that it is blue) does not give you any information about an other event (it is even).

Note that mutually exclusivity (disjoint events) and independence are not the same. Mutually exclusive events are not independent, and independent events cannot be mutually exclusive. Events are mutually exclusive if P (A and B) = 0.

So our independent events - blue and even - are not mutually exclusive, they can occur at the same time. Put another way, because events can't happen at the same time (disjoint or mutually exclusive), they can't be independent.

So if we take two mutually exclusive events - say the probability of a poker chip being both green (A) and blue (B) - we can check for the three parts of independence as:

1. Probability of event A is P(A) = 3/10 (.3)
2. Probability of event B is P(B) = 2/10 (.2)
3. Probability of their intersection is P(AB) = 0 which is not the same as P(A) P(B) (which is 0.06)

As P (AB) = 0 i.e. they are mutually exclusive they are dependent - knowing one i.e. the chip blue DOES give you information about whether the other event occured - you know it is not green, so the probability of being green goes from 30% before being told, to 0% after being told it is blue.

When events are mutually exclusive, when you know one thing is true the likelihood of the otehr being true becomes zero

For more than two events, we define independence the same way - the events are independent if the probability of their intersection is equal to the product of their probabilities.

2.1.8 Conditional Probability

What if knowing one event has occured tells us something about the probability that another event occured? How can we 'update' our knowledge in the event that the first event has occured?

The probability of A conditional on B is denoted as P(A|B). So the probability of A conditional on B, P(A|B), is P(AB)/ P(B), assuming P(B) > 0. We don't condition on an event if the probability of an event is 0%.

So in effect, by knowing one event has occured, it changes or re-defines our numerator for event B AND it is changing or re-defining our denominator - the part of the sample space which is now relevant - of event B.

There is a relationship between indepdence and conditional probability. Suppose A and B are independent and P(B) > 0. Then,

P(A|B) = P(AB)/P(B) = P(A)P(B) (as they are indepdent this is the same as P(AB)) / P(B) = P(A) (we cancel out P(B) from the previous)

or simply

P(A|B) = P(AB)/P(B) = P(A)P(B)/P(B) = P(A)

2.1.9 Conditional Probability in American Presidential Politics

If candidates for Republican nomination had the following probabilities - these might be obtained from looking at betting markets

Trump $P(A_1) = .4$
Cruz $P(A_2) = .3$
Rubio $P(A_3) = .2$
Carson $P(A_4) = .1$

How can we compute the probability of a Republican win for the presidency or P(W) i.e. the general election?

Conditional on winning the nomination, the candidates have following probabilities of winning the general election:

Trump $P(W|A_1) = .25$
Cruz $P(W|A_2) = .2$
Rubio $P(W|A_3) = .6$
Carson $P(W|A_4) = .4$

The probability of a Republic win is equal to the probability of the intersection between a Republican win and the sample space.

The sample space is the union between the four events A1 through A4. A1 through A4 are mutually exclusive and exhaustive events and therefore form a partition.

In terms of notation, we therefore have:

$P(W) = P(WS)$

$= P(W(A_1 \cup A_2 \cup A_3 \cup A_4))$ because A1-A4 are mutually exclusive and exhaustive sets, a partition
$= P(WA_1 \cup WA_2 \cup WA_3 \cup WA_4)$
$= P(WA_1) + P(WA_2) + P(WA_3) + P(WA_4)$
$= P(W|A_1)P(A_1) + P(W|A_2)P(A_2) + P(W|A_3)P(A_3) + P(W|A_4)P(A_4)$

So $P(W) = .4x.25 + .3x.2 + .2x.6 + .1x.4 = .32$

2.1.10 Bayes' Theorem

So far, we have seen that the probability of the intersection between A and B is equal to the Probability of B conditional on A times the probability of A:

- $P(AB) = P(B|A)P(A) = P(A|B)P(B)$
- provided $P(A) > 0$ and $P(B) > 0$ i.e. both A and B have positive probabilities
- so we can write $P(A|B) = P(B|A)P(A)/P(B)$

We also saw a slightly more complicated version of this, where the probability of B is the probability of B conditional on A times the probability of A, plus the probability of B conditional on A complement times the probability of A complement (note we saw this, albeit with more compliments, when looking at the Conditional Probability in American Presidential Politics section)

- $P(B) = P(B|A)P(A) + P(B|Ac)P(Ac)$
- $P(A|B) = P(B|A)P(A)/\{P(B|A)P(A) + P(B|Ac)P(Ac)\}$

C is compliment, and we can do this since A and Ac are partitions of the sample space S.

A pregnant woman lives in an area where the Zika virus is fairly rare - 1 in 1000 people have it. Still, she's concerned, so she gets tested. There is a good but not perfect test for the virus—it gives a positive reading with probability .99 if the person has the virus and a positive reading with probability .05 if the person does not. Her reading is positive. How concerned should we be?

$P(Z) = .001$ (unconditional probability of having Zika) $P(Zc) = .999$ (999 people don't have it) $P(+|Z) = .99$ (probability of having a positive test result, conditional on having the zika virus - there is a 1% change of a false negative) $P(+|Zc) = .05$ (probability of having a positive result if you don't have the virus is 5% - false positive rate) $P(Z|+) = P(+|Z)P(Z)/\{P(+|Z)P(Z) + P(+|Zc)P(Zc)\}$ - Bayes theorem = .019 - less than 2% probability

So the introduction of our new data results in us updating our probability based on the imperfect test, but it doesn't get updated by much as it still possible it's wrong and the prevelance rate of the zika virus is rare.

Example 2

Assume that the probability of having a rare condition is 1%. It is possible to test for the condition, but the test is imperfect. If you have the condition, there is an 85% chance that you will test positive. If you do not have the condition, there is a 5% chance that you will test positive. Call the condition C, so that P(C) = 0.01, and call a positive test t+, so that p(t+|C) = 0.85.

What is the probability p(t+) that you test positive for the condition?

So the Probability of having the condition is P(C) 0.01 * P(t+|C) = .85 which is the probability at a test you will test positive = 0.0085 + P(Cc) * P(t+|Cc) = 0.99 * 0.05 = 0.0495 = 0.058

Suppose that you tested positive for the condition. What is the probability that you truly have the underlying condition?

P(C) = .01 (unconditional probability of having condition) P(Cc) = .99 (99 people don't have it) P(t+|C) = .85 (probability of having a positive test result, conditional on having the condition) P(t+|Cc) = .05 (probability of having a positive result if you don't have the virus is 5% - false positive rate) P(C|+) = P(t+|C)P(C)/{P(t+|C)P(C) + P(t+|Cc)P(Cc)} - Bayes theorem = 0.0085 / {0.0085 + 0.0495} = .15 - around than 15% probability

Suppose that a new test is developed that is more accurate. Now, the probability of testing positive if you have the condition is 94%, and the chance of testing positive if you do not have the condition is only 4%. Now, what is the probability p(t+) that you test positive for the condition?

So the Probability of having the condition is P(C) 0.01 * P(t+|C) = .94 which is the probability at a test you will test positive = 0.0094 + P(Cc) * P(t+|Cc) = 0.99 * 0.04 = 0.0396 = 0.049

Suppose that you tested positive for the condition. What is the probability that you truly have the underlying condition?

P(C) = .01 (unconditional probability of having condition)
P(Cc) = .99 (99 people don't have it)
P(t+|C) = .94 (probability of having a positive test result, conditional on having the condition)
P(t+|Cc) = .04 (probability of having a positive result if you don't have the virus is 5% - false positive rate)
P(C|+) = P(t+|C)P(C)/{P(t+|C)P(C) + P(t+|Cc)P(Cc)} - Bayes theorem
= 0.0094 / {0.0094 + 0.0396}
= .19 - around than 15% probability

Suppose that there is an 80% chance you will be invited to a dinner party on a Friday or Saturday evening. In contrast, there is only a 50% chance that you will be invited to a dinner party on one of the other nights of the week. Suppose that you know that you've been invited to a dinner party tonight, but have forgotten which day of the week it is. Once you know that you've been invited to a dinner party, what is the chance that it is either Friday or Saturday? (Please round your answer to 2 decimal places. For example, if the correct answer is 0.6724, please input 0.67.)

Hint: Using the notation of Zika question, Let Z:= { Fri, Sat} and Z^c = { M,T,W,Th,Sun}. Let "+" denote invitation. You are given Pr("+"| Z) = 0.8 and Pr("+"| Z^c) = 0.5. We want to compute Pr(Z | "+")

P(Z) = .286 (unconditional probability of it being Friday or Saturday)
P(Zc) = .714 (the other 5 days of the week)
P(+|Z) = .8
P(+|Zc) = .5
P(Z|+) = P(+|Z)P(Z)/{P(+|Z)P(Z) + P(+|Zc)P(Zc)} - Bayes theorem
= .389 - around 40% probability

2.2 Random Variables, Distributions and Joint Distributions

A *random variable* is a real-valued function whose domain is the sample space - it goes from the sample space to the real line.

A probability goes from the set of all subsets of the sample space in to the unit interval e.g. [0,1] between zero and 1

A random variable goes from the sample space to the real line and it has some numerial charecteristics of the sample space we are interested in.

The probability that something exists induces a distribution of the random variable, they are not the same.

There are two types of random variable:

- Discrete - one that can take on only a - finite or infinite - countably number of values
- Continous - a random variable that can take on any value in some interval, bounded or unbounded, of the real line

Discrete random variables can be approximated using a continous random variable, so we typically just use continous. Most of the example we have seen so far in this section have dealt with discreet random variables.

2.2.1 Probability Functions of Random Variables

For discrete random variables, we often start with a verbal description, calculate probabilities for each value of the random variable, and then write down a function or draw a graph describing those probabilities for different values of the random variable. This is called a probability function (PF). We saw one of these before in the hypogeometric and binomial, when looking at the pizza toppings.

Note that:

- The term probability density function is used to draw attention to the fact that we are discussing a continuous random variable.
- The term probability mass function is used to draw attention to the fact that we are discussing a discrete random variable.
- The term probability function - or sometimes just the term “distribution” - is used when we are speaking in more general terms, when we’re discussing both “flavors” of probability function or the distinction between the two types of probability functions/random variables doesn’t matter.

Hypergeometric (pizza topping) random variable:

1 Verbal description - Let X be the number of vegetarian toppings I get on my pizza if I draw the Area Four toppings randomly (without replacement)

2 Calculation - We can calculate the probability that X = 0, 1, 2, and so forth, up to the maximum of 6 or n, whichever is smaller, using the formula from last time. Six is the maximum number of veg toppings available, n is the number of toppings chosen at random. If there are 0 toppings of a particular type, the result will be undefined, so we adjust 0! to be defined as just 1. Also, to be consistent with notation for the random variable, n1 from before now becomes x and since we only have two options, n2 now becomes n - x

$$P(xveg, n - xnon - veg) = \binom{6}{x}\binom{5}{n-x}\binom{11}{n}$$

3 If we then take an example, such as 3 veg toppings - n = 3 - we can calculate the probabilities for each n

P(X=0) = 6/99
P(X=1) = 36/99
P(X=2) = 45/99
P(X=3) = 12/99

And we can represent the probability function graphically, with points (aka point mass) then add vertical lines under each point to the axis to make it easier to read e.g.

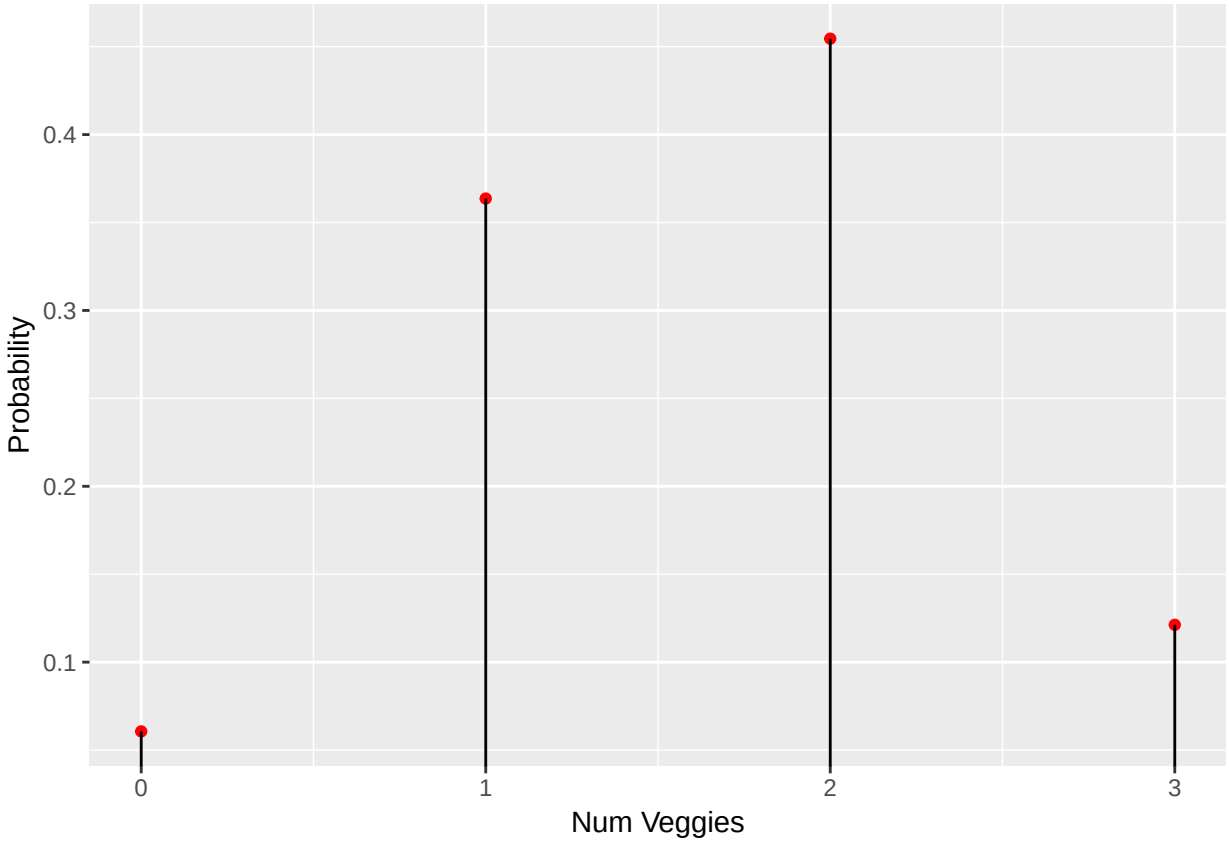
```
library(ggplot2)

veggie_choices = 6
meat_choices = 5
num_toppings = 3
veggie_received = 0:num_toppings
v = dhyper(x = veggie_received,
           m = veggie_choices,
           n = meat_choices,
           k = num_toppings)

for (i in 1:length(v)) {
  print(paste0("Probability of ",
               i-1,
               " veggie toppings is: ",
               round(v[i], 3)))
}
```

```
## [1] "Probability of 0 veggie toppings is: 0.061"
## [1] "Probability of 1 veggie toppings is: 0.364"
## [1] "Probability of 2 veggie toppings is: 0.455"
## [1] "Probability of 3 veggie toppings is: 0.121"
```

```
ggplot(mapping = aes(x = 0:3, y = v)) +
      geom_point(color = 'red') +
      labs(x = 'Num Veggies', y = 'Probability') +
      geom_segment(xend = 0:3, yend=0)
```



2.2.2 The Hypergeometric Distribution

We can represent this in a more general way using notation. We say that X has a "hypergeometric distribution with parameters N, K, & n," denoted X ~ H(N,K,n). Where

- N = Total number of toppings
- K = Total number of veg toppings
- n = The number we choose

Its Probability Function (PF) is defined similar to before, however we add a note for which values of x that there is positive probability. We should, if being fully formal, also add a final part which states it is 0 otherwise. If this is not explicit, as shown below, in terms of the zero otherwise, we can assume this to be the case.

$$fx(x) = \binom{K}{x}\binom{N-K}{n-x}\binom{N}{n}$$

$$where\ x = max(0, n+K-N), ..., min(n, K)$$

The hypergeometric distribution describes the number of number of "realized successes" (in a given sample - represented as x) in n trials where you're sampling without replacement from a sample of size N, whose initial probability of success was K/N.

The function provides the probability of X (number of successful outcomes / number of possible outcomes in the sample space).

2.2.3 Steph Curry Shooting example

If Steph has a probability of making 44% of any shot taken and therefore 56% chance of missing, we can use the binomial formula to calculate the probability of making n shots out of 6 possible shots as follows.

*For more information see the Binomial Coefficient

X has a "binomial distribution with parameters n & p," denoted $X \sim B(n, p)$. Its PF is

$$fx(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad where\ x = 0, 1, ...n$$

The binomial distribution describes the number of "successes" in n trials where the trials are independent and the probability of success in each is p.

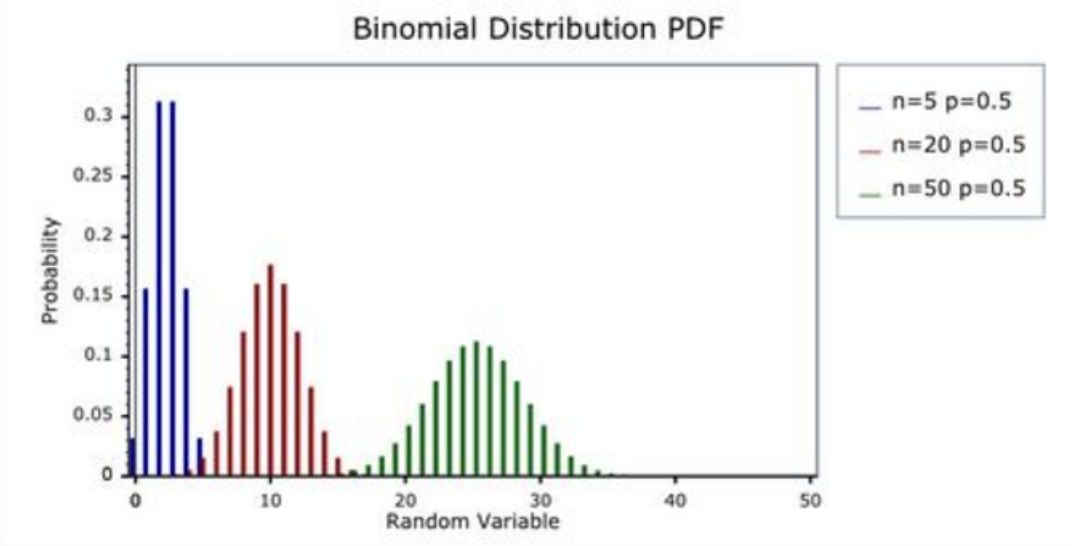
So plugging in our example we get

$$fx(x) = \binom{6}{x} .44^x (.56)^{6-x}$$

Which yields:

P(X=0) = .03
P(X=1) = .15
P(X=2) = .29
P(X=3) = .30
P(X=4) = .18
P(X=5) = .06
P(X=6) = .01

As the number of n increases, if p = 50% (a symetric distribution), the distribution would begin to look like a normal distribution.



In another example, suppose that you will take 3 penalty kicks in a row. The likelihood of making each penalty kick is ¾ or 75%. What is the probability that you will score 2 (and only 2) of the 3 penalty kicks?

$$fx(x) = \binom{3}{x} .75^x (.25)^{3-x}$$

P(X=0) = .02
P(X=1) = .14
P(X=2) = .42 <- this is the answer
P(X=3) = .42

2.2.4 Properties of the Probability Distribution

So for a general probability function, we have some broad properties:

- $0 <= f_x(x_i) <= 1$ which is to say the value of any probability function is going to be between 0 and 1
- ${}_i f_x(x_i) = 1$ if you sum up over all of the possible values it will sum to 1
- $P(A) = P(XcA) = {}_A f_x(x_i)$ which is to say if you want the probability over a set of values of x, you just sum up the individual values for each item in the set

For a continous random variable, we rarely start with a verbal description. Instead, we typically have a density that describes the probability that the random variable is in various regions. The density, or probability density function (PDF) is the continuous compliment to the discrete PF. The PF (discret) and PDF (continous) are similar but not exactly the same.

A random variable X is continuous if there exists a nonnegative function f_X_ such that for any interval A c R as follows. We tend to speak about a region, that A is in a region of the real line (R), the probability that X is in A is equal to the integral over that region A of the PDF.

$$P(XcA) = \int_A f_X(x)dx$$

2.2.5 Discrete versus Continuous Random Variables

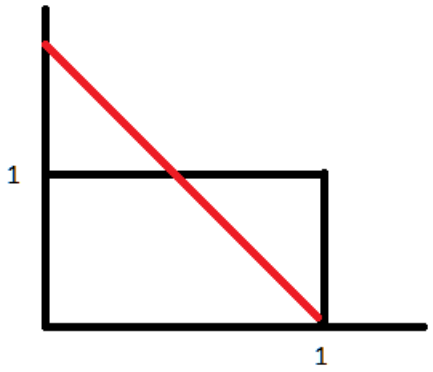
Just like the discrete probability function, the probability density function for a continous random variable has certain properties, these are:

- $0 <= f_X(x)$ which is to say it is non-negative. A PDF, unlike a PF, may have a region where the PDF is greater than 1

- $\int f_X(x) = 1$ the PF sums to one for each discreet part, the PDF integrates to one (as it is continous)
- $P(A) = P(a <= X <= b) = \int_A f_X(x)dx$ with discreet we sum for the values of interest, with PDF we integrate over the region (between a and b)

The value at a particular X for a PDF is equal to zero, if X is a continous random variable.

In terms of point 1, we can think of it like the image below. The area represented by the black lines will integrate to 1, however the area under the red line will have a region (top left, above the line) where it will integrate to more than 1.



In terms of any particular point (value of X) being zero, this is because at any single point on a continuous random variable, it is infinitely small and the integral of a single point is always zero. Or from wikipedia (!):

> Suppose a species of bacteria typically lives 4 to 6 hours. What is the probability that a bacterium lives exactly 5 hours? The answer is 0%. A lot of bacteria live for approximately 5 hours, but there is no chance that any given bacterium dies at exactly 5.0000000000… hours.
> Instead one might ask: What is the probability that the bacterium dies between 5 hours and 5.01 hours? Suppose the answer is 0.02 (i.e., 2%). Next: What is the probability that the bacterium dies between 5 hours and 5.001 hours? The answer should be about 0.002, since this time interval is one-tenth as long as the previous. The probability that the bacterium dies between 5 hours and 5.0001 hours should be about 0.0002, and so on.
> In these three examples, the ratio (probability of dying during an interval) / (duration of the interval) is approximately constant, and equal to 2 per hour (or 2 hour$^{\wedge}-1$). For example, there is 0.02 probability of dying in the 0.01-hour interval between 5 and 5.01 hours, and (0.02 probability / 0.01 hours) = 2 hour$^{\wedge}-1$. This quantity 2 hour$^{\wedge}-1$ is called the probability density for dying at around 5 hours.
> Therefore, in response to the question "What is the probability that the bacterium dies at 5 hours?", a literally correct but unhelpful answer is "0", but a better answer can be written as (2 hour$^{\wedge}-1$)dt. This is the probability that the bacterium dies within a small (infinitesimal) window of time around 5 hours, where dt is the duration of this window.
> For example, the probability that it lives longer than 5 hours, but shorter than (5 hours + 1 nanosecond), is 2 hour$^{\wedge}-1$ (1 nanosecond) $6\times10-13$ (using the unit conversion 3.6×1012 nanoseconds = 1 hour).
> There is a probability density function f with f(5hours)=2hour$^{\wedge}-1$. The integral of f over any window of time (not only infinitesimal windows but also large windows) is the probability that the bacterium dies in that window.

2.2.6 A Note on Terminology and the Uniform Distribution

In both text books and online, there can be differences in both the the terminology for random variables and the notation used. As noted earlier, we tend to use PMF (or just PF) for discreet RVs, PDF for continous and sometimes just PF or the "distribution of a random variable" when talking more broadly about both. It's perhaps best just to try and be consistent.

There can also be mixed random variables. One example might be if our measuring technology is such that we cannot measure past a certain point, so our variable is continous up to that point, then all values beyond that point get truncated or grouped (it is a probability mass and is discreet) to that maximal value.

There are some random variables which are simply uniform. We call such a random variable X "uniform with parameters and b" denoted as $X \sim U[a, b]$. In such a situation, the probability of X is defined in such a way

that the probability of X belonging to any subinterval of X is proportional to the length of the subinterval. Graphically, it looks like a box and similar to the last image above (the black line).

To calculate the probability of some interval [c,d] in [a,b] you integrate a/(b-a) over that region, or as the PDF is flat, we can just use (d-c)/(b-a). e.g. if we have a random variable that is uniformly-distributed from 3 to 8. What is the probability that the random variable takes on a value less than or equal to 7?

(7-3) / (8-3) = 4/5 = 0.8 or 80%

2.2.7 The Cumulative Distribution Function

Both discreet and continous random variables can be expressed in the form of a continous random variable (CDF) which takes on a value between 0 and 1. It is defined as:

$$f_X(x) = P(X <= x)$$

Also note that $lim_{x-}\ FX(x) = 0$
$lim_x\ FX(x) = 1$

So the CDF will start at zero, it may have flat parts, but it will never decrease. And as we go to the limit (x approaches infinity) then the CDF will be equal to 1.

Given a CDF it would be possible to recover the PDF or PF for continous or discreet distribution respectively.

- So if you want to get the CDF for a continuous random variable at a particular point, then you integrate the PDF up to that point: $FX(x) = P(X <= x) = -fX(x)dx$
- If you have the CDF and you want to get the PDF i.e. you want to recover the PDF from it, then you take the derivative: $FX(x) = \frac{dF(x)}{dx} = fX(x)$

2.2.8 Joint Distributions

A lot of what we do in data analysis is gather repeated observations from joint distributions of random variables, for instance, rainfall and crop growth. Such a two way joint distribution is called a bivariate distribution.

More formally we say

If X and Y are continuous random variables defined on the same sample space S, then the joint probability density function of X&Y, f_X_Y(x,y), is the surface such that for any region A of the xy-plane - note this is similar to the generalisation of the PDF but as a generalisation of the bivariate distribution:

$$P((X,Y)cA) = AfXY(x,y)dxdy$$

Like the PDF from before, it will integrate over the entire area to 1, and the probability at any one particular point is equal to zero.

*Video on single and double integration including limit

2.2.9 Joint Distribution Example

If you develop a headache you might decide to take tablets, one may be paracetamol and the other ibuprofen. If X is the effective period of ipbuprofen and Y that of paracetamol, then

$$fXY(x,y) = {}^2exp{-}(x+y)\ \ for x,y >= 0$$