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

James Solomon-Rounce

2018-09-21

Contents

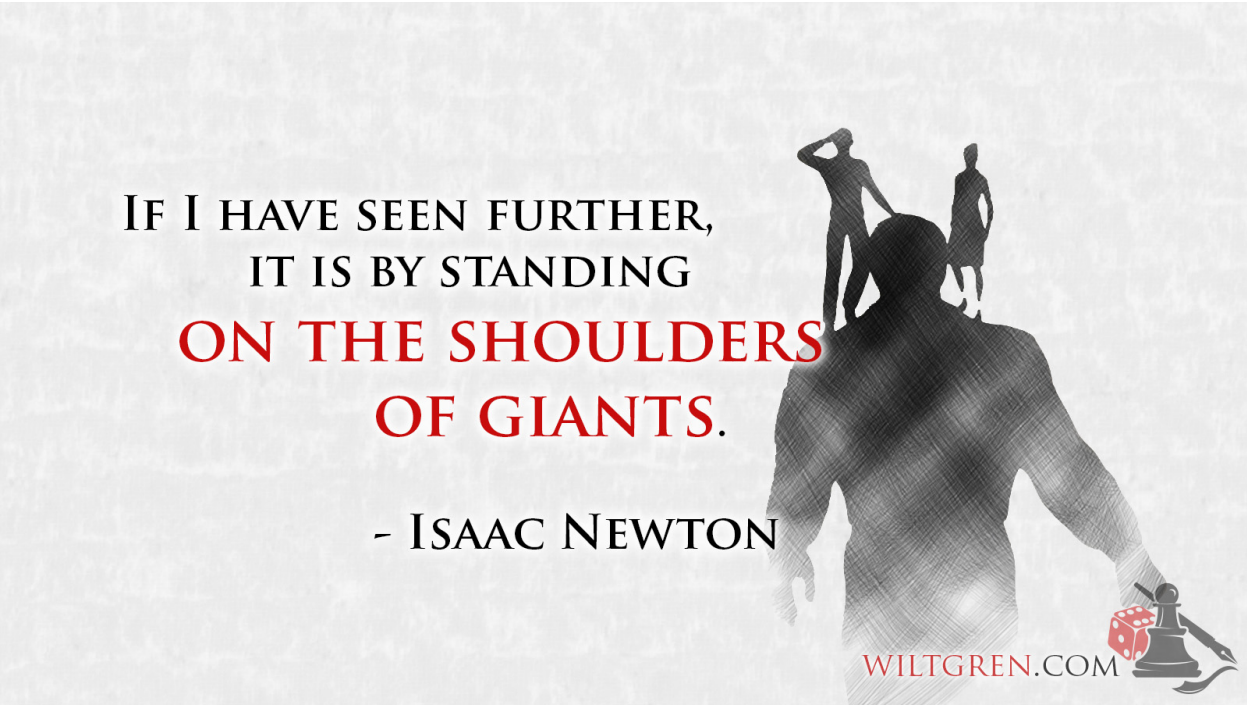
Preface

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:

- Module 2 Slides

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:

$P(W) = P(WS)$

$= P(W(A1 \text{ U } A2 \text{ U } A3 \text{ U } A4))$ because A1-A4 are mutually exclusive and exhaustive sets, a partition
$= P(WA\ 1 \text{ U } WA2 \text{ U } WA3 \text{ U } WA4)$
$= P(WA_1) + P(WA_2) + P(WA_3) + P(WA_4)$
$= P(W|A_1)P(A_1) + P(W|A_2)P(A_2) + P(W|A_3)P(A_3) + P(W|A_4)P(A_4)$

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