DSCI354-451 Foundation : Inference, EDA, DSCI Process(CWRU, Pitt, UCF, UTRGV)

Profs: R. H. French, L. S. Bruckman, P. Leu, K. Davis, S. Cirlos

TAs: W. Oltjen, K. Hernandez, M. Li, M. Li, D. Colvin

04 October, 2022

Contents

5.2.2.1 Class Readings, Assignments, Syllabus Topics

5.2.2.1.1 Reading, Lab Exercises, SemProjects

- Readings:
  - For today: R4DS 4-6
    - https://r4ds.had.co.nz/
  - For next class: R4DS 7-8
- Laboratory Exercises:
  - LE3 : Due Thursday October 6th
  - LE4 : Given out Friday October 7th
- Office Hours: (Class Canvas Calendar for Zoom Link)
  - Wednesday @ 4:00 PM to 5:00 PM, Will Oltjen
  - Saturday @ 3:00 PM to 4:00 PM, Kristen Hernandez
  - **Office Hours are on Zoom, and recorded**
- Semester Projects
  - DSCI 451 Students Biweekly Update 1 Due
  - DSCI 451 Students
    - Next **Report Out #1 is Due Friday September 30th**
  - All DSCI 351/351M/451 Students:
    - **Peer Grading of Report Out #1 is Due October 11th, 2022**
  - Exams
    - MidTerm: **Tuesday October 18th, in class or remote, 11:30 - 12:45 PM**
    - Final: Monday December 19, 2022, 12:00PM - 3:00PM, Nord 356 or remote

5.2.2.2 Textbooks

- Peng: R Programming for Data Science
- Peng: Exploratory Data Analysis with R
- Open Intro Stats, v4
- Wickham: R for Data Science
- Hastie: Intro to Statistical Learning with R, 2nd Ed.

Introduction to R and Data Science

- For R, Coding, Inferential Statistics
  - Peng: R Programming for Data Science
  - Peng: Exploratory Data Analysis with R

Textbooks for this class

- OIS = Diez, Barr, Çetinkaya-Runde: Open Intro Stat v4
- R4DS = Wickham, Grolemund: R for Data Science

Textbooks for DSCI353/353M/453, And in your Repo now

- ISLR2 = James, Witten, Hastie, Tibshirani: Intro to Statistical Learning with R 2nd Ed.
- ESL = Trevor Hastie, Tibshirani, Friedman: Elements of Statistical Learning
- DLwR = Chollet, Allaire: Deep Learning with R

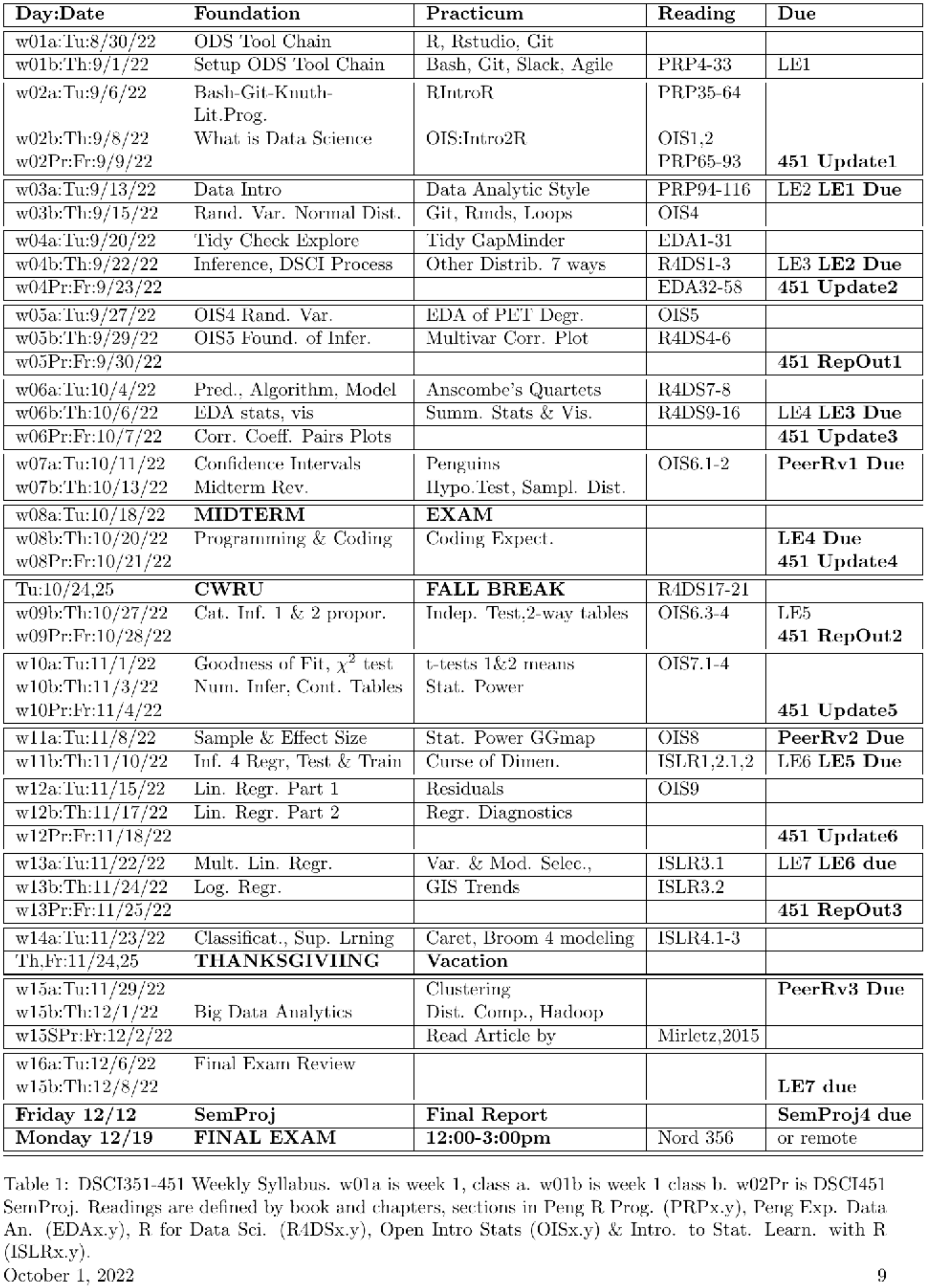
Magazine Articles about Deep Learning

- DL1 to DL13 are "Deep Learning" articles in 3-readings/2-articles/

5.2.2.3 Syllabus

5.2.2.4 Tidyverse Cheatsheets, Functions and Reading Your Code

- Look at the Tidyverse Cheatsheet



Day:Date	Foundation	Practicum	Reading	Due
w01a:Tu:8/30/22	ODS Tool Chain	R, Rstudio, Git		
w01b:Th:9/1/22	Setup ODS Tool Chain	Bash, Git, Slack, Agile	PRP4-33	LE1
w02a:Tu:9/6/22	Bash-Git-Knuth-Lit.Prog.	RIntroR	PRP35-64	
w02b:Th:9/8/22	What is Data Science	OIS:Intro2R	OIS1,2	
w02Pr:Fr:9/9/22			PRP65-93	**451 Update1**
w03a:Tu:9/13/22	Data Intro	Data Analytic Style	PRP94-116	LE2 **LE1 Due**
w03b:Th:9/15/22	Rand. Var. Normal Dist.	Git, Rmds, Loops	OIS4	
w04a:Tu:9/20/22	Tidy Check Explore	Tidy GapMinder	EDA1-31	
w04b:Th:9/22/22	Inference, DSCI Process	Other Distrib. 7 ways	R4DS1-3	LE3 **LE2 Due**
w04Pr:Fr:9/23/22			EDA32-58	**451 Update2**
w05a:Tu:9/27/22	OIS4 Rand. Var.	EDA of PET Degr.	OIS5	
w05b:Th:9/29/22	OIS5 Found. of Infer.	Multivar Corr. Plot	R4DS4-6	
w05Pr:Fr:9/30/22				**451 RepOut1**
w06a:Tu:10/4/22	Pred., Algorithm, Model	Anscombe's Quartets	R4DS7-8	
w06b:Th:10/6/22	EDA stats, vis	Summ. Stats & Vis.	R4DS9-16	LE4 **LE3 Due**
w06Pr:Fr:10/7/22	Corr. Coeff. Pairs Plots			**451 Update3**
w07a:Tu:10/11/22	Confidence Intervals	Penguins	OIS6.1-2	**PeerRv1 Due**
w07b:Th:10/13/22	Midterm Rev.	Hypo.Test, Sampl. Dist.		
w08a:Tu:10/18/22	**MIDTERM**	**EXAM**		
w08b:Th:10/20/22	Programming & Coding	Coding Expect.		**LE4 Due**
w08Pr:Fr:10/21/22				**451 Update4**
Tu:10/24,25	**CWRU**	**FALL BREAK**	R4DS17-21	
w09b:Th:10/27/22	Cat. Inf. 1 & 2 propor.	Indep. Test,2-way tables	OIS6.3-4	LE5
w09Pr:Fr:10/28/22				**451 RepOut2**
w10a:Tu:11/1/22	Goodness of Fit, χ^2 test	t-tests 1&2 means	OIS7.1-4	
w10b:Th:11/3/22	Num. Infer, Cont. Tables	Stat. Power		
w10Pr:Fr:11/4/22				**451 Update5**
w11a:Tu:11/8/22	Sample & Effect Size	Stat. Power GGmap	OIS8	**PeerRv2 Due**
w11b:Th:11/10/22	Inf. 4 Regr, Test & Train	Curse of Dimen.	ISLR1,2.1,2	LE6 **LE5 Due**
w12a:Tu:11/15/22	Lin. Regr. Part 1	Residuals	OIS9	
w12b:Th:11/17/22	Lin. Regr. Part 2	Regr. Diagnostics		
w12Pr:Fr:11/18/22				**451 Update6**
w13a:Tu:11/22/22	Mult. Lin. Regr.	Var. & Mod. Selec.,	ISLR3.1	LE7 **LE6 due**
w13b:Th:11/24/22	Log. Regr.	GIS Trends	ISLR3.2	
w13Pr:Fr:11/25/22				**451 RepOut3**
w14a:Tu:11/23/22	Classificat., Sup. Lrning	Caret, Broom 4 modeling	ISLR4.1-3	
Th,Fr:11/24,25	**THANKSGIVIING**	**Vacation**		
w15a:Tu:11/29/22		Clustering		**PeerRv3 Due**
w15b:Th:12/1/22	Big Data Analytics	Dist. Comp., Hadoop		
w15SPr:Fr:12/2/22		Read Article by	Mirletz,2015	
w16a:Tu:12/6/22	Final Exam Review			
w15b:Th:12/8/22				**LE7 due**
Friday 12/12	**SemProj**	**Final Report**		**SemProj4 due**
Monday 12/19	**FINAL EXAM**	**12:00-3:00pm**	Nord 356	or remote

Table 1: DSCI351-451 Weekly Syllabus. w01a is week 1, class a. w01b is week 1 class b. w02Pr is DSCI451 SemProj. Readings are defined by book and chapters, sections in Peng R Prog. (PRPx.y), Peng Exp. Data An. (EDAx.y), R for Data Sci. (R4DSx.y), Open Intro Stats (OISx.y) & Intro. to Stat. Learn. with R (ISLRx.y).

October 1, 2022 9

Figure 1: DSCI351-351M-451 Syllabus

  - **Tidyverse For Beginners Cheatsheet**
    - In the Git/20s-dsci353-353m-453-prof/3-readings/3-CheatSheets/ folder
  - **Data Wrangling with dplyr and tidyr Cheatsheet**

Tidyverse Functions & Conventions

  - The pipe operator `%>%`
  - Use `dplyr::filter()` to subset data row-wise.
  - Use `dplyr::arrange()` to sort the observations in a data frame
  - Use `dplyr::mutate()` to update or create new columns of a data frame
  - Use `dplyr::summarize()` to turn many observations into a single data point
  - Use `dplyr::arrange()` to change the ordering of the rows of a data frame
  - Use `dplyr::select()` to choose variables from a tibble,
    - keeps only variables you mention
  - Use `dplyr::rename()` keeps all the variables and renames variables
    - rename(iris, petal_length = Petal.Length)
  - These can be combined using `dplyr::group_by()`
    - which lets you perform operations "by group".
  - The `%in%` matches conditions provided by a vector using the c() function
  - The **forcats** package has tidyverse functions
    - for factors (categorical variables)
  - The **readr** package has tidyverse functions
    - to read_…, melt_… col_…, parse_… data and objects

Reading Your Code: Whenever you see

- The assignment operator `<-`, think **"gets"**
- The pipe operator, `%>%`, think **"then"**

5.2.2.5 Statistical Inference, Exploratory Data Analysis, and the Data Science Process

5.2.2.5.1 Doing Data Science

- What to be thinking about as we do DSCI

5.2.2.5.2 Statistical Thinking in the Age of Big Data The Age of Big Data, Steve Lohr, The New York Times

- Also in 3-readings/2-articles in your class repo

Big Data is a vague term, used loosely, if often, these days. But put simply, the catchall phrase means three things.

- First, it is a bundle of technologies.
- Second, it is a potential revolution in measurement.
- And third, it is a point of view, or philosophy,
  - about how decisions will be
  - and perhaps should be
  - made in the future.

Data Science is a practical mixture of Statistics, Coding, Linear Algebra,

- And Domain Knowledge

5.2.2.6 Frequentist vs. Bayesian Statistics

5.2.2.6.1 Frequentist Statistics has to do with

- Inferring the properties of the population
- From a sample taken from the population
- And it is based in
  - probability densities
  - and observed statistical frequency in a sample
- It is the most common type of inferential statistics

5.2.2.6.2 Bayesian Statistics focuses on

- Current observed results
  - referred to as priors
- And tries to infer what future observed results will be
  - Using the information already acquired from Priors
- It doesn't use the "Take a Sample from a Population" approach
- Popular for problems that are difficult using Frequentist Statistics

5.2.2.7 Lets get founded in statistical inference

- Different from descriptive statistics
- which is basically reductive
  - Calculating the mean and standard deviation
    - reduces the number of values in your dataset
    - and destroys information!

5.2.2.7.1 Statistical Inference

- The world we live in is complex, random, and uncertain.
- At the same time, it's one big data-generating machine.
- Data represents the traces of the real-world processes,
- and exactly which traces we gather are decided by our data collection or sampling method.
- You, the data scientist, the observer,
  - are turning the world into data,
  - and this is an utterly subjective, not objective, process.

There are two sources of randomness and uncertainty.

- the randomness and uncertainty underlying the process itself,
- and the uncertainty associated with your underlying data collection methods.

Once you have all this data, you have somehow captured the world, or certain traces of the world.

So you need a new idea, and that's to simplify those captured traces into something more comprehensible,

- to something that somehow captures it all in a much more concise way,
- and that something could be mathematical models or functions of the data,

These are known as statistical estimators.

This overall process of going

- from the world to the data,
- and then from the data back to the world,

Is the field of statistical inference.

5.2.2.8 Populations and Samples

- In classical statistical literature,
    - a distinction is made between the population and the sample.

5.2.2.8.1 Population

- The word population immediately makes of people
- But it can be any set of objects or units,
    - such as tweets or photographs or stars.

If we could measure the characteristics of all those objects,

- we'd have a complete set of observations,
- and the convention is to use N to represent
    - the total number of observations in the population.

5.2.2.8.2 Sample

- When we take a sample,
    - we take a subset of the units of size n
    - in order to examine the observations to draw conclusions
    - and make inferences about the population.

5.2.2.9 Populations and Samples of Big Data

- But, wait!
    - In the age of Big Data,
        * where we can record all users' actions all the time,
        * don't we observe everything?
    - Is there really still this notion of population and sample?
    - If we had all the email in the first place,
        * why would we need to take a sample?

Sampling solves some engineering challenges

- the focus on Hadoop to handle engineering and computational challenges - caused by too much data
- Overlooks sampling as a legitimate solution.
    - At Google, for example, software engineers, data scientists, and statisticians
    - sample all the time.

5.2.2.9.1 Bias

- Any inferences we make from that data should not be extended
- to draw conclusions about humans beyond those sets of users,
- or even those users for any particular day.
    - Example of tweets and hurricane Sandy

5.2.2.9.2 Sampling

- Let's rethink what the population and the sample are in various contexts.

In statistics we often model the relationship between

- a population and a sample
- with an underlying mathematical process.

So we make simplifying assumptions about

- the underlying truth,
- the mathematical structure, and shape
  - of the underlying generative process that created the data.

We observe only one particular realization

- of that generative process,
- which is that sample.

Sampling is beneficial to our thinking

- Since a sample of a population
- Will change with the next sampling we do
- So we intrinsically have the uncertainty
- And variability of sampling, too keep us honest!

5.2.2.10 Big Data Can Mean Big Assumptions

- "Big" is a moving target.

  - Big Data isn't a size.
    - i.e. 1 petabyte is big?
    - it can be, but maybe not
    - depends on the data

"Big" is when you can't fit it on one machine.

- Not a useful distinction
- I've been doing things since Cray1
- A good computer now is like the Cray1
- In our research group we use >200 computers
- Scaleable analytics is critical
- And number of computers is irrelevant
  - Except if you only demo things on your mac
  - That is not big data, that phone data

Big Data is a cultural phenomenon.

- It describes how much data is part of our lives, precipitated by accelerated advances in technology.
  - this has some validity

The 4 Vs:

- this seems to work
  - Volume, variety, velocity, and value.

5.2.2.10.1 N = All?

- Collecting and using a lot of data rather than small samples
- Accepting messiness in your data
- Giving up on knowing the causes + Not so good

Data is not objective

- Another way in which the assumption that N=ALL can matter
  - is that it often gets translated into the idea that data is objective.
- It is wrong to believe either that data is objective or that "data speaks,"
- and beware of people who say otherwise.

5.2.2.10.2 N = 1

- The other end of the spectrum from N=ALL, we have n = 1,
  - by which we mean a sample size of 1.
  - In the old days a sample size of 1 would be ridiculous;
  - you would never want to draw inferences about an entire population by looking at a single individual.

But the concept of n = 1 takes on new meaning in the age of Big Data,

- where for a single person, we actually can record tons of information about them,
- and in fact we might even sample from all the events or actions they took
  - (for example, phone calls or keystrokes)
- in order to make inferences about them.
- This is what user-level modeling is about.

5.2.2.11 Modeling

- Data "models" are schema on how to store data for computer scientists
- Data Scientists want statistical models or mathematical models

Who is following Andrew Gelman on twitter?

- Who knows who Andrew Gelman is?
- Who knows where he is?
- http://andrewgelman.com/
- Good discussion of current Statistical Topics
  - http://andrewgelman.com/2017/09/26/abandon-statistical-significance/

5.2.2.11.1 What is a Model

- A model is our attempt to understand and represent
  - the nature of reality through a particular lens,
  - be it architectural, biological, or mathematical.
- A model is an artificial construction
  - where all extraneous detail has been removed or abstracted.
  - Attention must always be paid to these abstracted details
  - after a model has been analyzed
    - to see what might have been overlooked.

5.2.2.11.2 Statistical Models and Greek and Latin Letters (Important)

- Notation in Statistics

In mathematical expressions, the convention is

- to use Greek letters for parameters
- and Latin letters for data.

So, for example, if you have two columns of data, x and y,

- and you think there's a linear relationship,
- you'd write down $y = \beta_0 + \beta_1 x + \epsilon$.
  - Where ϵ is a random error term
  - In this way, the equation represents the actual data points
  - Not the fitted line (which only approximately fits the data)

You don't know what β_0 and β_1 are in terms of actual numbers yet,

- so they're the parameters.

We do this notation in an .Rmd file

- Using LaTeX Math Mode for the symbols
- Using inline math mode
  - β_0 and β_1
- Or using normal math mode

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

5.2.2.11.3 How do you build a model

- One place to start is exploratory data analysis (EDA)
- This entails making plots and building intuition for your particular dataset.
- EDA helps out a lot,
- Another approach to modeling is trial and error and iteration.

5.2.2.11.4 Remember, it's always good to start simply.

- There is a trade-off in modeling between simple and accurate.
  - Simple models may be easier to interpret and understand.
  - Oftentimes the crude, simple model gets you 90% of the way there
  - and only takes a few hours to build and fit,
- whereas getting a more complex model
  - might take months and only get you to 92%.

5.2.2.12 Probability Distributions

- Probability distributions are the foundation of statistical models.

Back in the day, before computers,

- scientists observed real-world phenomena, took measurements,
- and noticed that certain mathematical shapes kept reappearing.

The classical example is the height of humans,

- following a normal distribution—a bell-shaped curve,
- also called a Gaussian distribution, named after Gauss.

Other common shapes have been named after their observers as well

- (e.g., the Poisson distribution and the Weibull distribution),
- while other shapes such as Gamma distributions or exponential distributions
- are named after associated mathematical objects.

Natural processes tend to generate measurements

- whose empirical shape could be approximated
- by mathematical functions with a few parameters
- that could be estimated from the data.

Figure 2-1 as an illustration of the various common shapes,

- Remember they only have names
- because someone observed them enough times to think they deserved names.
- There is actually an infinite number of possible distributions.

They are to be interpreted as assigning a probability

- to a subset of possible outcomes,
- and have corresponding functions.

For example, the normal distribution is written as:

$$N(x|\mu,\sigma) \sim \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Figure 2: the normal distribution

The parameter μ is the mean and median

- and controls where the distribution is centered
- (because this is a symmetric distribution),

and the parameter σ controls

- how spread out the distribution is.

Note the Greek letter parameters (estimators)

- and the Latin letters for data (variables)

This is the general functional form,

- but for specific real-world phenomenon,
- these parameters have actual numbers as values,
- which we can estimate from the data.

5.2.2.13 A bunch of distributions for different uses

5.2.2.13.1 Distribution of a single random variable $p(x)$

- A random variable denoted by x or y can be assumed to have
  - a corresponding probability distribution, $p(x)$,
  - which maps x to a positive real number.

In order to be a probability density function,

- we're restricted to the set of functions
- such that if we integrate p (x) to get the area under the curve,
- it is 1, so it can be interpreted as probability.

5.2.2.13.2 Joint distributions $p(x, y)$

- In addition to denoting distributions of single random variables
  - with functions of one variable,

we use multivariate functions called joint distributions

- to do the same thing for more than one random variable.

So in the case of two random variables, for example,

- we could denote our distribution by a function p (x,y),
- and it would take values in the plane
- and give us non-negative values.

In keeping with its interpretation as a probability,

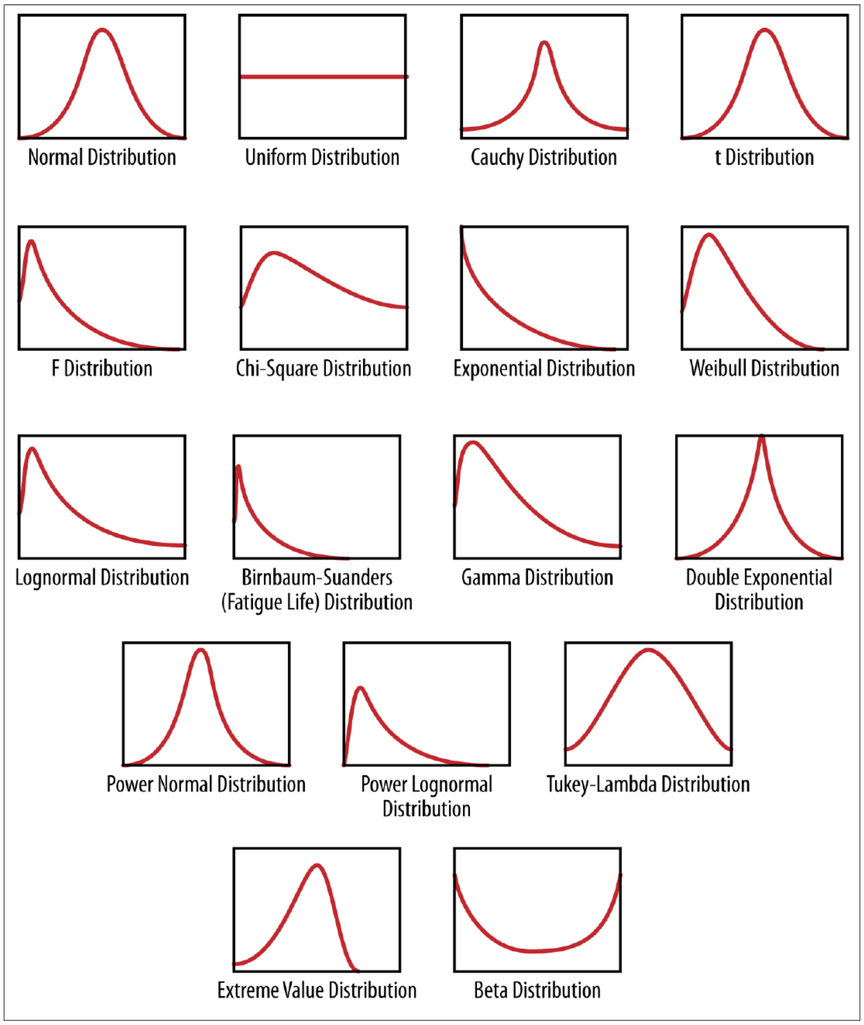


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

Figure 3: distributions

- its (double) integral over the whole plane would be 1.

5.2.2.13.3 Conditional Distributions $p(x|y)$

- A conditional distribution, $p(x|y)$
  - which is to be interpreted as
  - the density function of x
  - given a particular value of y.

When we're working with data,

- conditioning corresponds to subsetting.
- i.e. applying a condition to the dataset

Example

- suppose we have a set of user-level data for Amazon.com
- that lists for each user
  - the amount of money spent last month on Amazon,
  - whether the user is male or female,
  - and how many items they looked at before adding the first item to the shopping cart.

If we consider x to be the random variable that represents the amount of money spent,

- then we can look at the distribution of money spent across all users,
- and represent it as $p(x)$.

We can then take the subset of users

- who looked at more than five items before buying anything,
- and look at the distribution of money pent among these users.

Let y be the random variable that represents number of items looked at,

- then $p(x|y) > 5$ would be the corresponding conditional distribution.

Note a conditional distribution has the same properties

- as a regular distribution in that
- when we integrate it, it sums to 1 and has to take non-negative values.

5.2.2.14 Fitting a model

- Fitting a model means that
  - you estimate the parameters of the model
  - using the observed data.

You are using your data as evidence to help approximate the real-world mathematical process that generated the data.

Fitting the model often involves optimization methods and algorithms,

- such as maximum likelihood estimation,
- to help get the parameters.

In fact, when you estimate the parameters,

- they are actually estimators,
- meaning they themselves are functions of the data.

Once you fit the model,

- you actually can write it as y = 7.2 + 4.5x, for example,

- which means that your best guess is that
  - this equation or functional form
- expresses the relationship between your two variables,
- based on your assumption that the data followed a linear pattern.

Beware of overfitting

- The Bias-Variance Tradeoff
- https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff
  - More discussed in DSCI353-453
  - And in Introduction to Statistical Learning with R (ISLR book)

5.2.2.15 Exploratory Data Analysis

- "Exploratory data analysis" is an attitude,

  - a state of flexibility,
  - a willingness to look for those things that we believe are not there,
  - as well as those we believe to be there.
  - John Tukey

John Tukey, a mathematician at Bell Labs,

- developed exploratory data analysis in contrast to confirmatory data analysis,
- which concerns itself with modeling and hypotheses.

In EDA, there is no hypothesis and there is no model.

- The "exploratory" aspect means that
- your understanding of the problem you are solving, or might solve,
- is changing as you go.

The basic tools of EDA are

- plots,
- graphs and
- summary statistics.

Generally speaking, it's a method of systematically going through the data,

- plotting distributions of all variables (using box plots),
- plotting time series of data,
- transforming variables,
- looking at all pairwise relationships between variables
  - using scatterplot matrices,
- and generating summary statistics for all of them.

At the very least that would mean

- computing their mean,
- minimum,
- maximum,
- the upper and lower quartiles,
- and identifying outliers.

5.2.2.15.1 Philosophy of Exploratory Data Analysis

- Long before worrying about how to convince others,

  - you first have to understand what's happening yourself.
  - Andrew Gelman

5.2.2.16 The Data Science Process

- Let's put it all together into what we define as the data science process.

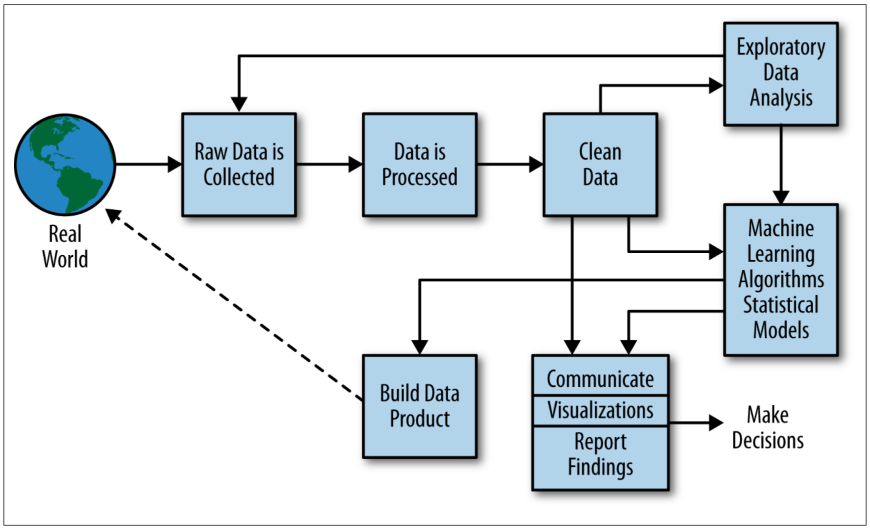


Figure 2-2. The data science process

Figure 4: the data science process

First we have the Real World.

Inside the Real World are lots of people busy at various activities.

We start with raw data

We want to process this to make it clean for analysis.

- So we build and use pipelines of data munging:
  - joining, scraping, wrangling,
  - or whatever you want to call it.
- To do this we use tools such as
  - Python,
  - shell scripts,
  - R,
  - or SQL,
  - or all of the above.

Once we have this clean dataset,

- we should be doing some kind of EDA.

In the course of doing EDA, we may realize

- that it isn't actually clean because of
  - duplicates, missing values, absurd outliers,
  - and data that wasn't actually logged or incorrectly logged.

If that's the case, we may have to go back

- to collect more data,
- or spend more time cleaning the dataset.

Next, we design our model to use some algorithm

- like k-nearest neighbor (k-NN), linear regression, Naive Bayes, or something else.

The model we choose depends on the type of problem we're trying to solve,

- The type of data science problem/question could be
- a classification problem,
- a prediction problem,
- or a basic description problem.

A Data Scientist's Role in This Process

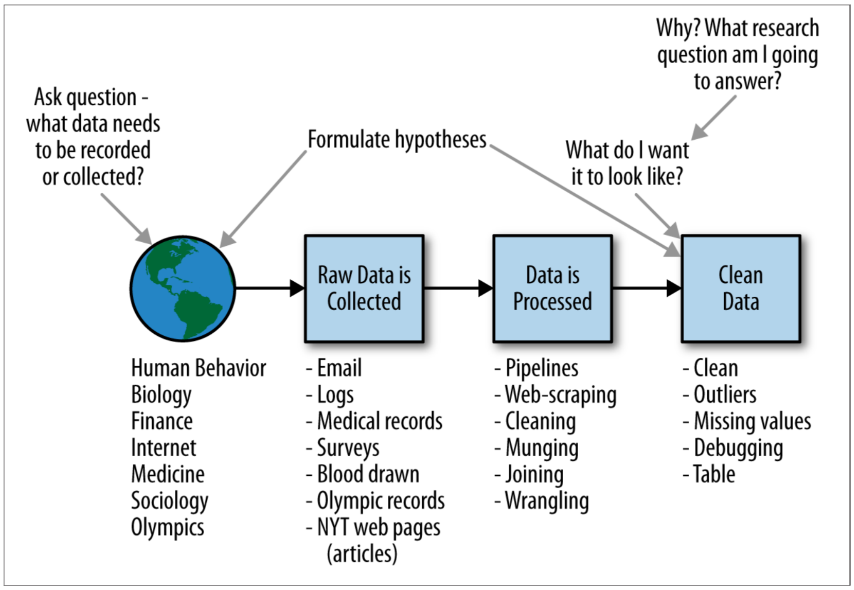


Figure 2-3. The data scientist is involved in every part of this process

Figure 5: the role of thedata scientist