Statistical Data Modelling

FIT5197

Monash University

About this Unit

Resources

1. Moodle contains
   - Unit Orientation, Assessments and Discussion Forums
   - links to Lecture Notes, recommended videos & readings, *Alexandria material*

Resources

1. Moodle contains
   - Unit Orientation, Assessments and Discussion Forums
   - links to Lecture Notes, recommended videos & readings, *Alexandria material*
2. S. Ross's *Probability and Statistics for Engineers and Scientists* 2014 (5th edition)
   - PDFs available from library, with answers to questions!
   - substantial parts of (but not all of) Chapters 1-11 and 15

Resources

1. Moodle contains
   - Unit Orientation, Assessments and Discussion Forums
   - links to Lecture Notes, recommended videos & readings, *Alexandria material*
2. S. Ross's *Probability and Statistics for Engineers and Scientists* 2014 (5th edition)
   - PDFs available from library, with answers to questions!
   - substantial parts of (but not all of) Chapters 1-11 and 15
3. *probability cheat sheet*, print out now in colour!

Resources

1. Moodle contains
   - Unit Orientation, Assessments and Discussion Forums
   - links to Lecture Notes, recommended videos & readings, *Alexandria material*
2. S. Ross's *Probability and Statistics for Engineers and Scientists* 2014 (5th edition)
   - PDFs available from library, with answers to questions!
   - substantial parts of (but not all of) Chapters 1-11 and 15
3. *probability cheat sheet*, print out now in colour!
4. using R in Jupyter Notebook. Alternatives:
   - install RStudio, *https://www.rstudio.com*, on your own computer
   - or use *https://jupyterhub.erc.monash.edu/*
   - or use *MoVE*
   - *Short R Reference Card* and *Base R Cheat Sheet*, print out now in colour!

Getting Started

1. Check in Moodle for Week 1
   - *FIT5197M1: Introduction to Modelling for Data Science* in Alexandria
   - videos

Getting Started

1. Check in Moodle for Week 1
   - *FIT5197M1: Introduction to Modelling for Data Science* in Alexandria
   - videos
2. Tutorial first week is learning R
   - material in Alexandria
   - use R in Jupyter Notebook or JupyterHub or RStudio or MoVE
   - install R in Jupyter Notebook on your own computer for later assignments

Getting Started

1. Check in Moodle for Week 1
   - *FIT5197M1: Introduction to Modelling for Data Science* in Alexandria
   - videos
2. Tutorial first week is learning R
   - material in Alexandria
   - use R in Jupyter Notebook or JupyterHub or RStudio or MoVE
   - install R in Jupyter Notebook on your own computer for later assignments
3. How these classes are run:
   - watch videos & read background material between classes
   - bring a device to lectures to participate in quizzes
   - prepare for tutes

Getting Started

1. Check in Moodle for Week 1
   - *FIT5197M1: Introduction to Modelling for Data Science* in Alexandria
   - videos
2. Tutorial first week is learning R
   - material in Alexandria
   - use R in Jupyter Notebook or JupyterHub or RStudio or MoVE
   - install R in Jupyter Notebook on your own computer for later assignments
3. How these classes are run:
   - watch videos & read background material between classes
   - bring a device to lectures to participate in quizzes
   - prepare for tutes
4. Want to learn more yourself?
   - work through Ross chapters 1-10; answers to questions available
   - in particular, work through the proofs he gives to better understand the why's

Assessment

	Week due	Content	Percent
Assign. 1	1a:4,1b:5,1c:7, 1d:8,1e:9,1f:12,	Hand written MCQ and SAQ (6 sub-parts)	6 x 2.5% per sub-part = 15%
Assign. 2	6	R program stats	20%
Assign. 3	10	R program model	15%
Exam	TBD	MCQ and SAQ	50%

- Exam is closed book but you have the formula sheet included, and can bring in non-programmable calculator.
- Assignments 1 (each of the 6 sub-parts) due back marked after 1 week.
- Assignments 2 and 3 due back marked after 2 weeks.
- No R in exam.

Academic Integrity

Monash University is committed to upholding standards of academic integrity and honesty. Monash students are therefore required to:

- undertake studies and research responsibly and with honesty and integrity;
- ensure that academic work is in no way falsified;
- seek permission to use the work of others, where required;
- acknowledge appropriately the work of others; and
- take reasonable steps to ensure that other students are unable to copy or misuse their work.

see *Student Academic Integrity Policy*

Contacts

Need help?

1. ask questions during tutorials and lectures
   - *please* interrupt me with questions!
   - am happy to work through proofs in Ross

Contacts

Need help?

1. ask questions during tutorials and lectures
   - *please* interrupt me with questions!
   - am happy to work through proofs in Ross
2. check for relevant Discussions Forum on Moodle
   - note in particular the "Assessments" discussion threads
   - but do NOT post your solutions to assignments ;-)
3. attend the consultation hour of the tutors or the lecturer
   - consultation hours in Moodle
4. send email to tutor or lecturer
5. contact Levin Kuhlmann (levin.kuhlmann@monash.edu) for special consideration

Content of the Unit

- technical overview of basic principles of probability and data analysis
- exposure to some common analytic models
- basic experience with R

Content of the Unit

- technical overview of basic principles of probability and data analysis
- exposure to some common analytic models
- basic experience with R

- students wanting more in depth machine learning after FIT5197:
  - work through the relevant chapters of Ross fully
  - I'll present various additional pointers during class
  - learn to use and learn from Wikipedia

Motivation for the Unit

- technologies in data science are constantly evolving

Motivation for the Unit

- technologies in data science are constantly evolving
- one thing that will never change is the importance of probability, statistics and modelling
  - despite what the popular press may tell you, deep neural networks also builds on these principles

Motivation for the Unit

- technologies in data science are constantly evolving
- one thing that will never change is the importance of probability, statistics and modelling
  - despite what the popular press may tell you, deep neural networks also builds on these principles
- understand this and you can then read the important text books and pick up on the new technologies for machine learning, etc.
  - consider *"gradient boosting machines"*: look at Wikipedia or Medium.com entry
  - Google's *Tensor Flow Probability talk* at AI Conference by O'Reilly.

Unit Schedule: Modules

Module	Week	Content	Ross
1.	1	Introduction to modelling for data science and to R	1,2
2.	2	Probabilities	3
	3	Expectations	4
	4	Distributions	5
3.	5	Statistical inference	6&7
	6	Hypothesis testing	7&8
4.	7	Dependence and linear regression	9
	8	classification and clustering	
5.	9	Comparing means	10
	10	Random number generation and simulation	
6.	11	Validation and complexity	15
	12	Modelling	

Any Other Questions?

... before we get started?

Any Other Questions?

... before we get started?

... otherwise, use the discussion board!

FIT5197 Statistical Data Modelling

Module 1

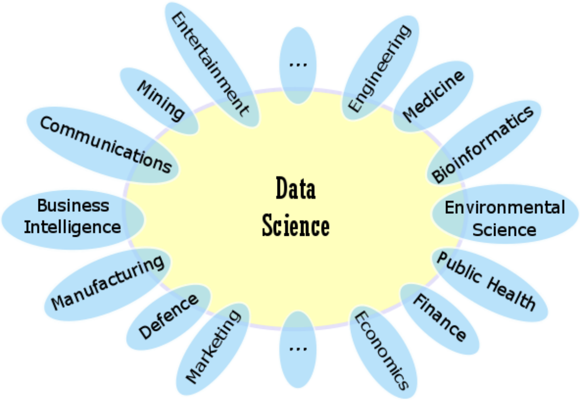
Introduction to modelling for data science and to R

2020 Lecture 1

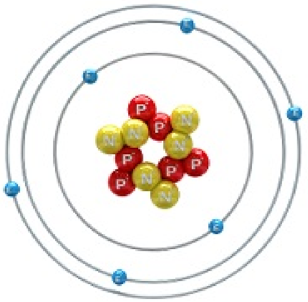
Monash University

Introduction to Modelling (ePub sections 1.2, 1.3, 1.4, 1.5,1.7, Ross Ch. 1)

Data Science Views

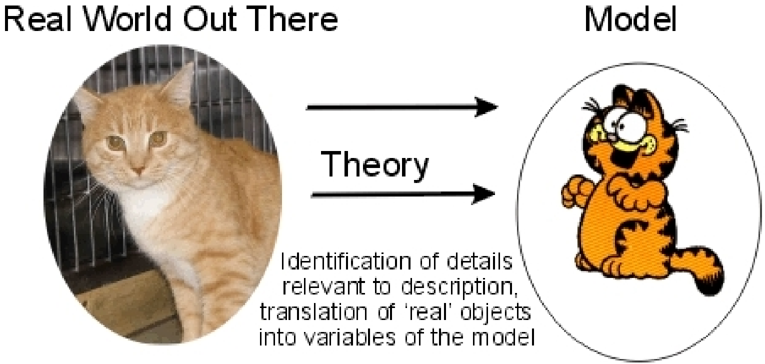


What Are Models?



- representation of a real world problem/object/process
- allows explanation, analysis and inference
- for us, usually a probabilistic model

Modelling

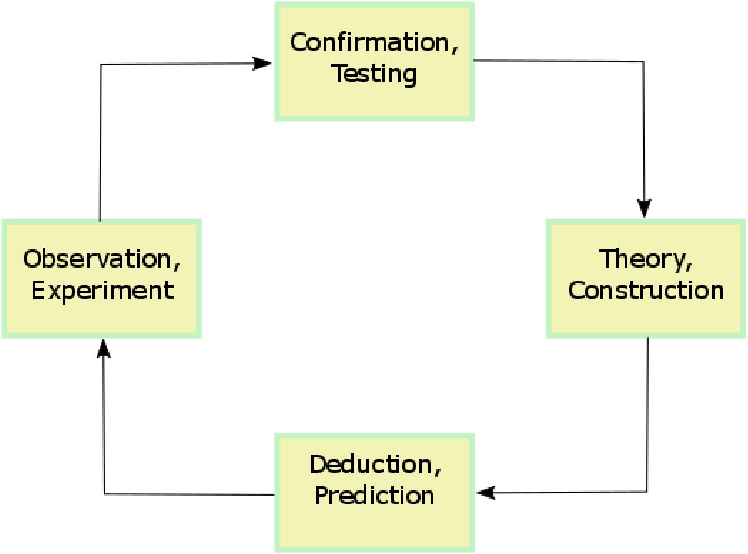


from *the BackReaction blog by Sabine Hossenfelder*

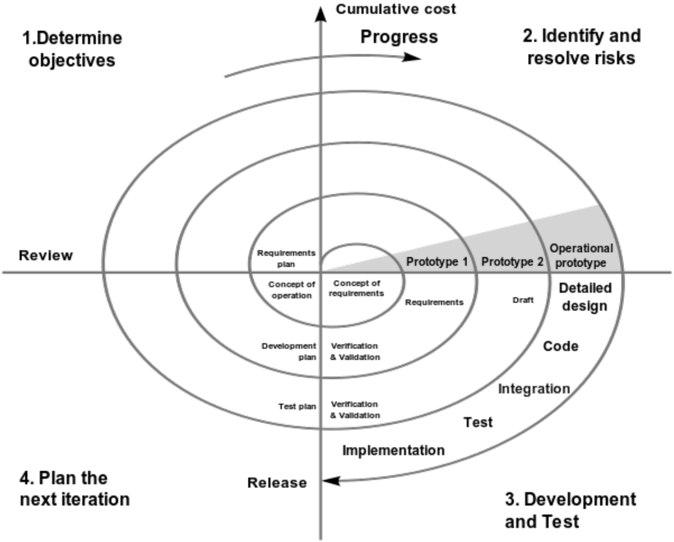
George Box (renowned statistician)

"Essentially, all models are wrong, but some are useful"

What is the Scientific Method?



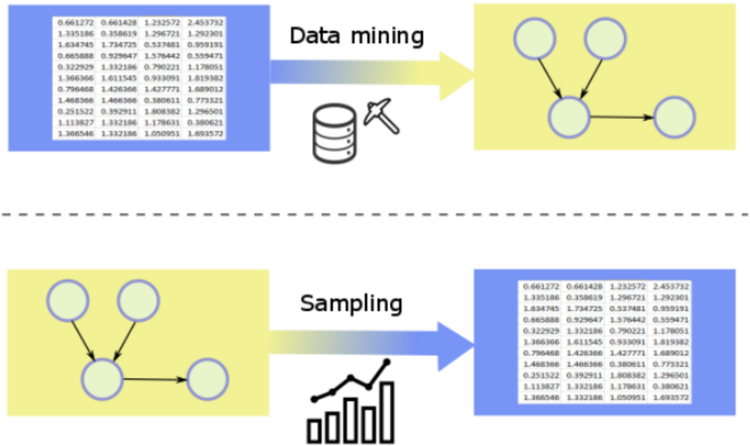
How to Build Complex Systems



Boehm's spiral model (1988)

continue prototyping until is reasonable

Data to Models and Back



Concepts for Data

Quantitative data: numerical values, always given to a fixed precision, including vectors, etc.

Qualitative data: categorical values, including structured data

Object: thing about which data is measured

Population: set of all objects of interest

Sample: data from a subset of the population

Inference: estimate properties of the population based on properties of the sample

Prediction: predict/guess of properties given a new set of sample

Inference/Prediction Examples

Book-seller: how many copies of this popular book will I sell this month?

Electicity company: what's the maximum kWhs needed at any one time to meet consumer demand in Clayton over summer?

Oncologist: how does this drug affect female patients with stage 3 bowel cancer who are over 60 years of age?

Hospital administrator: what sorts of patient cohorts are there for cardiology patients in the Intensive Care Unit so I can organise staff and treatment?

Inference/Prediction Examples

Book-seller: how many copies of this popular book will I sell this month?

Electicity company: what's the maximum kWhs needed at any one time to meet consumer demand in Clayton over summer?

Oncologist: how does this drug affect female patients with stage 3 bowel cancer who are over 60 years of age?

Hospital administrator: what sorts of patient cohorts are there for cardiology patients in the Intensive Care Unit so I can organise staff and treatment?

Warning: the traditional scientific method only tests very specific questions; in general in data analysis we can ask broader questions.

Where Does the Data Come From?

- To do inference we need a sample.
  - But where does it come from?

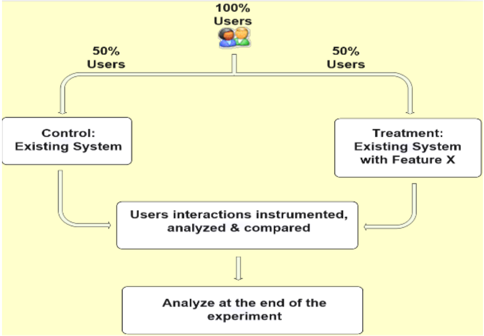
Where Does the Data Come From?

- To do inference we need a sample.
  - But where does it come from?
- *Controlled Experiments*: gather data that best supports inference about the questions under consideration.
  - *A/B testing*
  - *Randomised control trial (RCT)*
  - can be used to assess cause and effect

Where Does the Data Come From?

- To do inference we need a sample.
  - But where does it come from?
- *Controlled Experiments*: gather data that best supports inference about the questions under consideration.
  - *A/B testing*
  - *Randomised control trial (RCT)*
  - can be used to assess cause and effect
- Use observational data or survey data.
  - see *convenience sampling*
  - see *participation bias*
  - see *observational study*
  - ...

Controlled Experiments



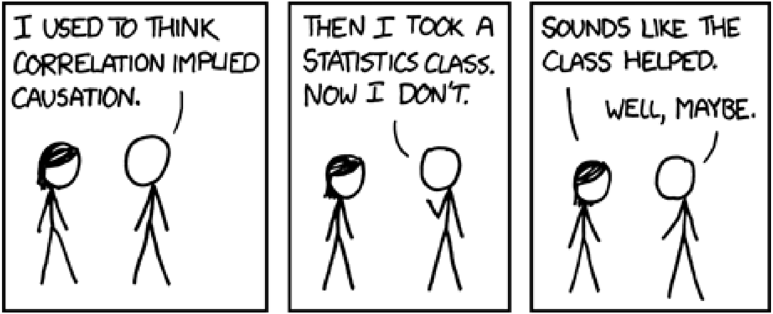
- Random allocation of treatment and control group
- Blinding: participants, administers, outcome accessors or data analyst were blinded of the allocation
- Attrition: loss to follow up

Observational studies

- *Case-control study*: compares those who with and without outcome of interest, and looks back retrospectively to identify risk factors for the outcome.
- *Cross-sectional study*: using data collection from a population at one specific point in time
- *Longitudinal study*: correlational research study that involves repeated observations of the same variables over long periods of time (i.e. *Cohort study*)
  - Longitudinal data
  - Time series data
  - Survival data

Causal Inference

Association or Correlation $\neq$ Causation



https://xkcd.com/552/

Causal Inference

Criteria for causal relationship by Sir Austin Bradford Hill

- Strength: strong associations more likely to be causal, but weak association does not rule out a causal connection
- Consistency: reproducibility
- Temporality: The effect has to occur after the cause
- Dose resonse relationship
- Plausibility: plausible mechanism between cause and effect
- Specificity, Coherence, Experiment and Analogy

Making causal inferences requires judgment about quality and quantity of evidence. Strong causal evidence: Systematic literature review and RCT

Data Problems

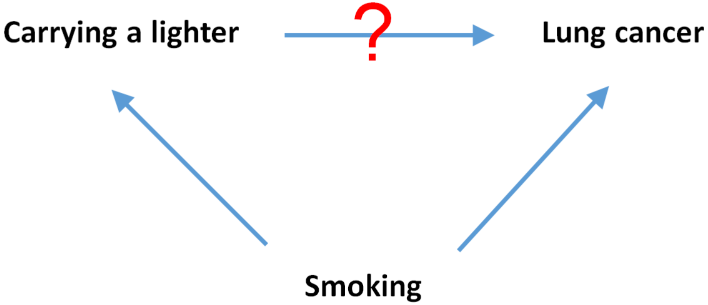
- Missing data
- Confounders
- Bias
- Outliers
- ...

Missing Data

- **Missing completely at random (MCAR)**: the missing data are just a random subset of the data
  - Complete cases analysis - normal approach, unbiased and conservative
  - Single value imputation - underestimation of standard errors
- **Missing at random (MAR)**: "missingness" is random given (i.e. conditional on) observed information
  - Complete cases analysis - biased
  - Single value imputation - underestimation of standard errors
  - Advanced methods: multiple imputation (MI), EM etc.
- **Missing not at random (NMAR)**:responders differ from non-responders, even after accounting for the observed information:
  - imputation methods such as MI and EM will not work
  - investigate missing mechanisms...

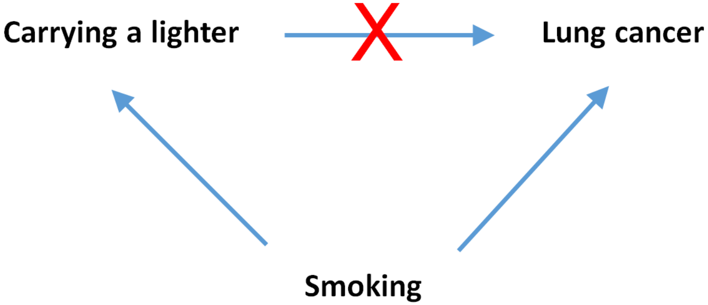
Dealing with Confounders

Confounding variable (confounders): variable that influences both the dependent variable and independent variable causing a spurious association.



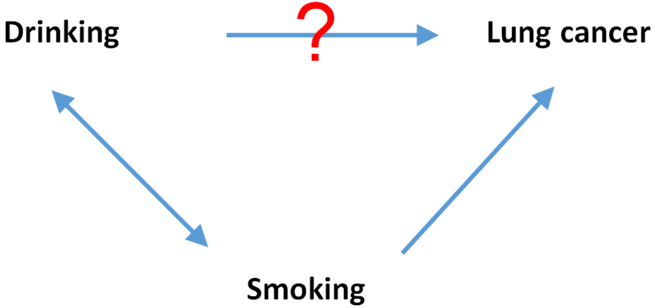
Dealing with Confounders

Confounding variable (confounders): variable that influences both the dependent variable and independent variable causing a spurious association.



Dealing with Confounders

Confounding variable (confounders): variable that influences both the dependent variable and independent variable causing a spurious association.



- *stratified sampling*
- or justification (include as a covariance)

Bias

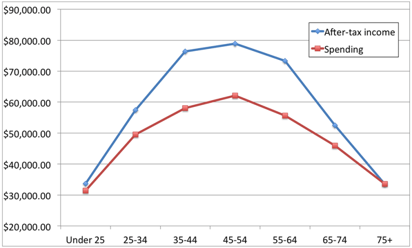
Bias: influence in the collection/analyses of data which results in systematically inaccurate estimates of population values.

- Selection or participation bias: members of a population are not selected uniformly randomly.
  - *1936 US presidential election (The Literary Digest poll)*
- Instrument bias: systematic bias in measurements
  - there is a *true NASA story* along these lines!
  - but they don't mention the data censoring done by NASA that caused the issue to *not be discovered* by NASA instead
- Statistical or inductive bias: bias inherent in a statistical or data mining tool

Statistical or Inductive Bias

When the model restricts the class of functions/fits one allows:

- predicting income from age using a linear model
  - there will be a negative bias at the end when the person retires
- predicting income from age using a piecewise linear model
  - you can make small linear pieces to fit the “true” function



from

Yes, Your Spending Will (Probably)
Decrease in Retirement

Outliers

Outliers are data points that lie well beyond the bulk of samples for a variable on one or more dimensions. Outliers be:

- legitimate data points
- due to measurement error
- from a heavy-tailed distribution

How to deal with outliers?

- delete
- correct the measurement error
- keep in the analysis

The choice of how to deal with an outlier should depend on the cause, type of analysis and the research question

Concepts for Modelling

- How do we infer properties of the entire population from a small sample?

Concepts for Modelling

- How do we infer properties of the entire population from a small sample?
- What sorts of models would we use for the population?

Concepts for Modelling

- How do we infer properties of the entire population from a small sample?
- What sorts of models would we use for the population?
- How would inference be done?

Concepts for Modelling

- How do we infer properties of the entire population from a small sample?
- What sorts of models would we use for the population?
- How would inference be done?
- How can we best predict when the new data arrived?

Questions for Data Science

Hypothesis Testing: main tool for all of empirical science

- largely recipe driven, (covered in Module 3)

Questions for Data Science

Hypothesis Testing: main tool for all of empirical science

- largely recipe driven, (covered in Module 3)

Modelling: build a “model” for a domain problem

- to be used for prediction, “understanding,” planning, (covered in Module 4)

Questions for Data Science

Hypothesis Testing: main tool for all of empirical science

- largely recipe driven, (covered in Module 3)

Modelling: build a “model” for a domain problem

- to be used for prediction, “understanding,” planning, (covered in Module 4)

Diagnostics: “debugging” a model or an algorithm

- is your model suitable for the problem?
- is the algorithm working?

Questions for Data Science

Hypothesis Testing: main tool for all of empirical science

- largely recipe driven, (covered in Module 3)

Modelling: build a "model" for a domain problem

- to be used for prediction, "understanding," planning, (covered in Module 4)

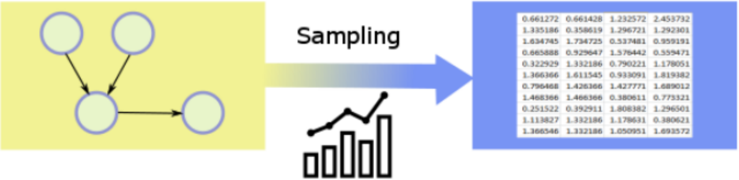
Diagnostics: "debugging" a model or an algorithm

- is your model suitable for the problem?
- is the algorithm working?

Algorithm Analysis and Design: techniques and issues

- don't have to "do", but should be aware of
- building up from parts,
- model "fitting", MLE, minimum cost
- bias-variance

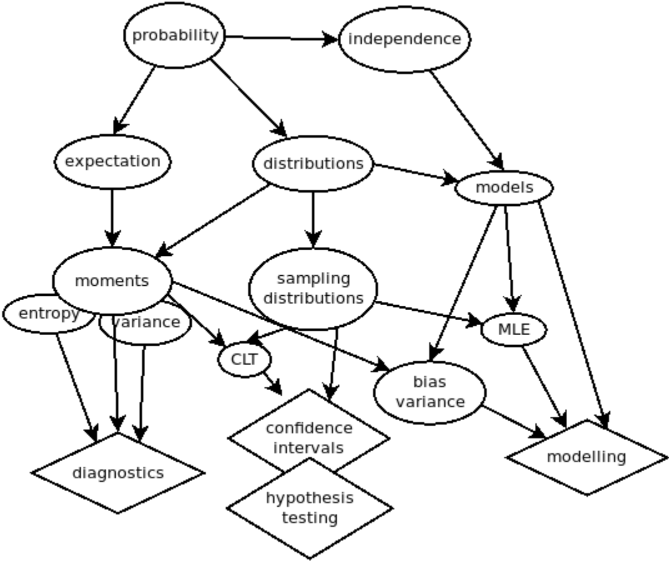
Generative Models



A generative model gives probabilities for every data point:

- **thus** probabilities for every sample
- **thus** lets us compare typical versus rare samples
- **thus** lets us test whether the model is reasonable or unreasonable for the sample
- is the basis for hypothesis testing (covered in Module 3)

Concept Map for This Unit



Exploratory Data Analysis (ePub section 1.7, Ross Ch. 2)

practical for this will be done in tutorial of Week 2

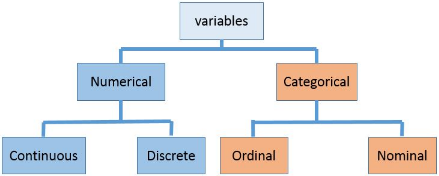
Data Matrix, Observations and Variables

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovatec
2	7129300520	20141013T000000	221900	3	1	1180	5650	1	0	0	3	7	1180	0	1955	
3	6414100192	20141209T000000	538000	3	2.25	2570	7242	2	0	0	3	7	2170	400	1951	199
4	5631500400	20150225T000000	180000	2	1	770	10000	1	0	0	3	6	770	0	1933	
5	2487200875	20141209T000000	604000	4	3	1960	5000	1	0	0	5	7	1050	910	1965	
6	1954400510	20150218T000000	510000	3	2	1680	8080	1	0	0	3	8	1680	0	1987	
7	7237550310	20140512T000000	1225000	4	4.5	5420	101930	1	0	0	3	11	3890	1530	2001	
8	1321400060	20140627T000000	257500	3	2.25	1715	6819	2	0	0	3	7	1715	0	1995	
9	2008000270	20150115T000000	291850	3	1.5	1060	9711	1	0	0	3	7	1060	0	1963	
10	2414600126	20150415T000000	229500	3	1	1780	7470	1	0	0	3	7	1050	730	1960	
11	3793500160	20150312T000000	323000	3	2.5	1890	6560	2	0	0	3	7	1890	0	2003	
12	1736800520	20150403T000000	662500	3	2.5	3560	9796	1	0	0	3	8	1860	1700	1965	
13	9212900260	20140527T000000	468000	2	1	1160	6000	1	0	0	4	7	860	300	1942	
14	114101516	20140528T000000	310000	3	1	1430	19901	1.5	0	0	4	7	1430	0	1927	
15	6054650070	20141007T000000	400000	3	1.75	1370	9680	1	0	0	4	7	1370	0	1977	
16	1175000570	20150312T000000	530000	5	2	1810	4850	1.5	0	0	3	7	1810	0	1900	

Variables

- a *variable* is any characteristic, number, or quantity that can be measured or counted
- there are different types of variables based on the ways they are studied, measured and represented:

- **Numerical** (quantitative)
  - **Continuous**
  - **Discrete**
- **Categorical** (qualitative)
  - **Ordinal**
  - **Nominal**



Variables, cont

- **Numerical** (quantitative)
  Have values that describe a measurable quantity as a number, like 'how many' or 'how much'. It is meaningful to do arithmetic.

  Continuous: Observations can take any value between two specified values.

  Discrete: Observations can take a value based on a count from a set of distinct values

- **Categorical** (qualitative)
  Have values that describe a 'quality' or 'characteristic' of a data unit, like 'what type' or 'which category'

  Ordinal: Observations can take a value that can be logically ordered or ranked

  Nominal: Observations can take a value that is not able to be organized in a logical sequence

Examples

- **Variables:**
  Age, sex, business income and expenses, country of birth, capital expenditure, class grades, eye colour, vehicle type
- **Continuous** (how much)
  Height, time, age, and temperature
- **Discrete** (how many)
  Number of registered cars, number of business locations, and number of children in a family
- **Ordinal** (has order)
  Academic grades (e.g., A, B, C), clothing size (e.g., small, medium, large, extra large), attitudes (e.g., strongly agree, agree, disagree, strongly disagree), dates
- **Nominal** (no order)
  Gender, business type, eye colour, religion and brand

Have a Look at iris Data

- The iris dataset has been used for classification
- Consists of 50 samples from each of three classes of iris flowers
- Use str() to compactly display the structure of an arbitrary R object

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

Have a Look at `iris` Data

- The first or last rows of data can be retrieved with `head()` or `tail()`
- We can also retrieve the values of a single column

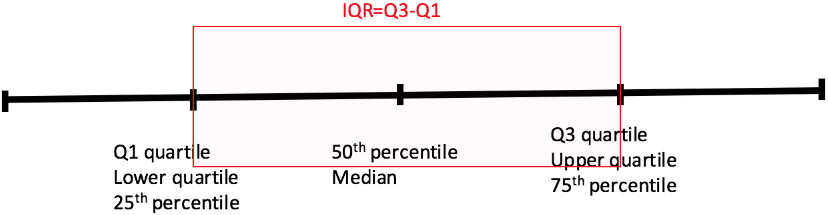
```
> head(iris)
> tail(iris)
> iris[1:5,]
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
> iris[1:10, "Sepal.Length"]
[1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9
> iris$Sepal.Length[1:10]
[1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9
```

Statistics on variables

- Measure of centre
  - Mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Median
    - Order the data
    - Find the mid-point or average of two mid-points
- Measure of spread
  - Variance, $\text{var} = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Standard deviation, $s_x = \sqrt{\text{var}}$
    - also called sd, s.d., std dev, etc.
  - Range $= \max_{i=1}^{n} x_i - \min_{i=1}^{n} x_i$
  - (inter-quartile range) IQR $= Q_3 - Q_1$
- Robust statistics: extreme observations have little effect
  - Median is more robust than mean
  - IQR more robust than range, variance and std
  - They are usually better for skewed data

Quartiles

- For sorted data
- Quartiles are 3 points that divide into 4 equal groups
- Each group is a quarter of data



```
> x <- c(0:10, 50)
> xm <- mean(x)
> xm
[1] 8.75

> median(1:4)
[1] 2.5
> median(1:5)
[1] 3
> median(c(1:3, 100, 1000))
[1] 3
> c(median(1:4), mean(1:4))
[1] 2.5 2.5
> c(median(1:5), mean(1:5))
[1] 3 3
> c(median(c(1:3, 100, 1000)), mean(c(1:3, 100, 1000)))
[1]   3.0 221.2
> var(1:20)
[1] 35
> sd(1:20)
[1] 5.91608
> sqrt(var(1:20))==sd(1:20)
[1] TRUE
```

```
> range(3:10)
[1]  3 10
> diff(range(3:10))
[1] 7
> IQR(c(3:10, 100, 1000))
[1] 4.5
> c(IQR(c(3:10, 100, 1000)), diff(range(c(3:10, 100, 1000))))
[1]   4.5 997.0
> boxplot(c(3:100, 150, 200))
```

Summary of Variables

- Distribution of every variable can be checked with function `summary()`
- it returns the min, max, mean, median, and the first (25%) and third (75%) quartiles.
- For factors (or categorical variables), it shows the frequency of every level.

```
> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
> summary(iris$Sepal.Length)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   5.100   5.800   5.843   6.400   7.900
> summary(iris)
  Sepal.Length    Sepal.Width    Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Summary of Variables

- The mean, median and range can also be obtained with functions with `mean()`, `median()` and `range()`
- Quartiles and percentiles are supported by function `quantile()`
- Use `var()` to check variance

```
> quantile(iris$Sepal.Length)
0% 25% 50% 75% 100%
4.3 5.1 5.8 6.4 7.9
> quantile(iris$Sepal.Length, c(.1, .3, .65))
10% 30% 65%
4.80 5.27 6.20
> var(iris$Sepal.Length)
[1] 0.6856935
```

Further Summary of Variables

```
> summary(cars)
     speed           dist       
 Min.   : 4.0   Min.   :  2.00  
 1st Qu.:12.0   1st Qu.: 26.00  
 Median :15.0   Median : 36.00  
 Mean   :15.4   Mean   : 42.98  
 3rd Qu.:19.0   3rd Qu.: 56.00  
 Max.   :25.0   Max.   :120.00  
> fivenum(cars$speed) # min, lower hinge, median, upper hinge, max.
[1]  4 12 15 19 25
> boxplot.stats(cars$speed) # Boxplot stats: fivenum (as above: min, lower hing
                # upper hinge, max), n, confidence interval (CI) of the median
$stats
[1]  4 12 15 19 25

$n
[1] 50

$conf
[1] 13.43588 16.56412

$out
numeric(0)
```

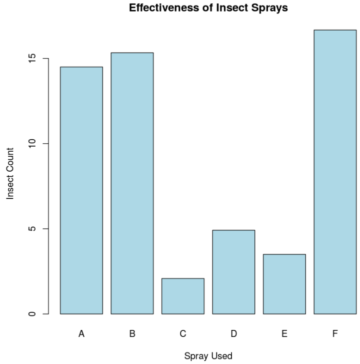
Graphical Representations

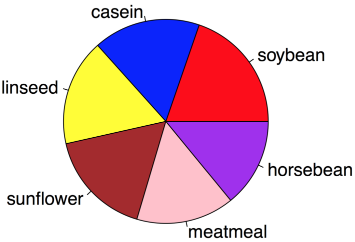
- To understand the properties of data
- To find possible pattern in data
- To guide us in choosing better and more suitable models
- To communicate the outcome with others

	Single variable	Two variables	Categorical
Numerical	Histogram Box plot	Scatter plot	Side-by-side box plo
Categorical	Pie chart Bar plot	segmented bar plot Mosaic plot	

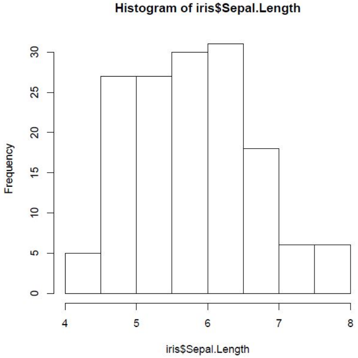
Data exploration choices

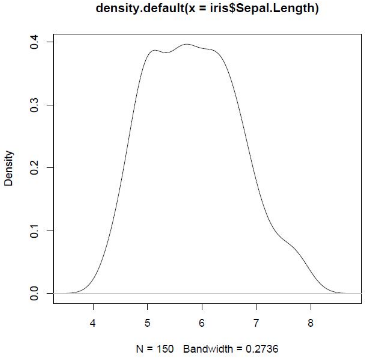
Single Categorical: Bar Chart and Pie Chart





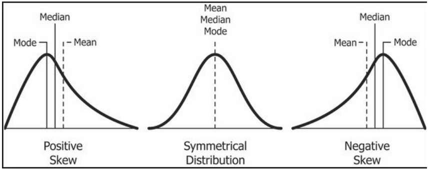
Explore Individual Variables

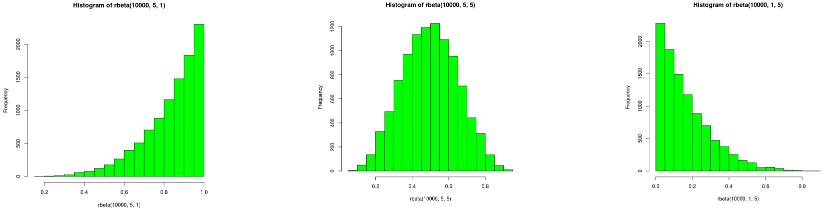




Visual Attributes of Distrib.s

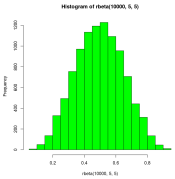
- Shape of skewness
  - Symmetric
  - Left skewed
  - Right Skewed

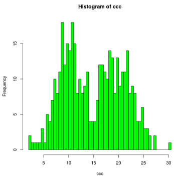


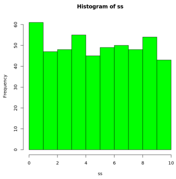


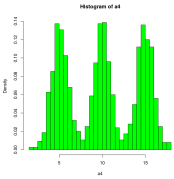
Visual Attributes of Distrib.s

- Modality
  - Unimodal
  - Bimodal
  - Uniform
  - Multimodal



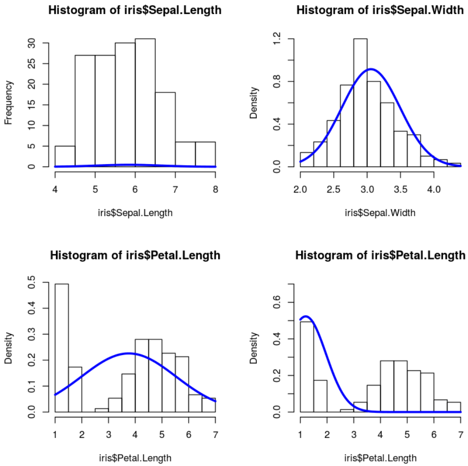






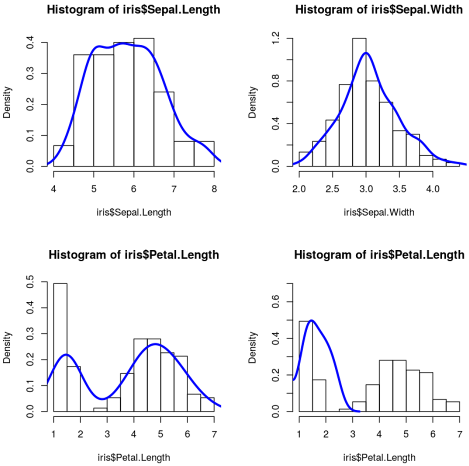
Histogram and Density

curve(dnorm(x, mean = mean(iris$Petal.Width), sd = sd(iris$Petal.Width)), add = TRUE)

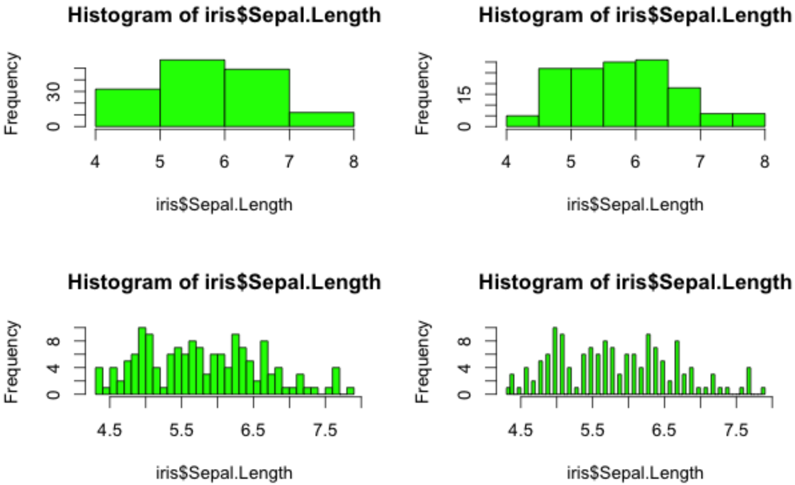


Histogram and Density

lines(density(iris$Petal.Width), col = "blue", lwd = 3)

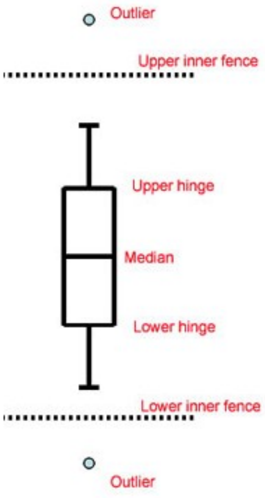


Different Number of Bars



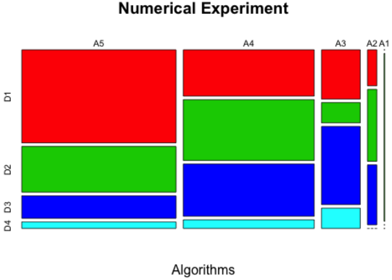
Boxplot for R

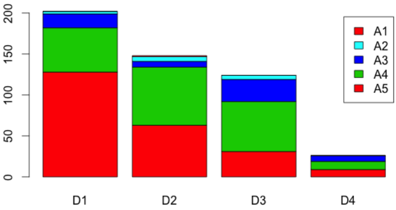
- introduced by *John W. Tukey*
- represent numerical data through quartiles (lower hinge is Q_1, upper is Q_3)
- whiskers indicating variability outside the upper and lower quartiles
- lower inner fence is $Q1 - 1.5 \times \mathrm{IQR}$
- upper inner fence is $Q3 + 1.5 \times \mathrm{IQR}$
- whiskers at min/max data values inside fences
- highlights outliers outside fences



(there are many versions of layouts for a boxplot, and the default in R (seems) to be the Tukey boxplot, see *boxplot types*)

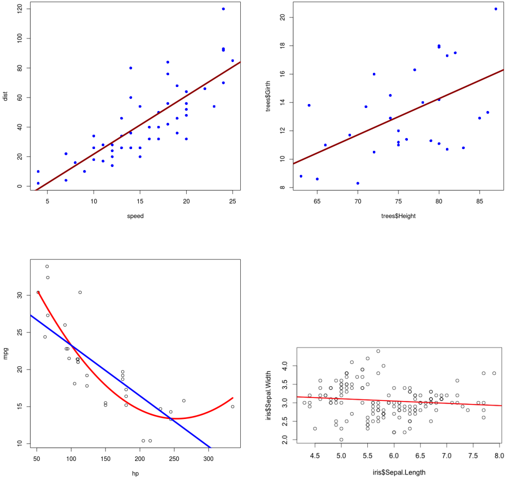
Segmented Bar Chart and Mosaic Plot





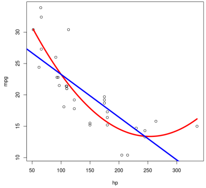
Relationship Between Two Numerical Variables

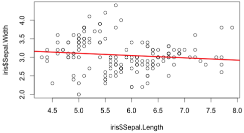
- Direction of relationship
  - Increasing
  - Decreasing
- Shape
  - Linear
  - Nonlinear
- Strength
  - Strong
  - Weak

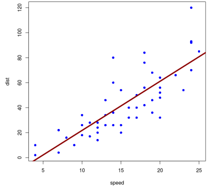


Covariance and Correlation

- *Covariance* of two random variables shows how they are related:
  - Positive covariance, then they are positively related
  - Negative covariance, then they are negatively related
- The *correlation coefficient* of two random variables is covariance divided by the product of their standard deviations:
  - it shows how the two random variable are linearly related
  - if close to 1, then they are positively linearly related
  - if close to -1, then they are negatively linearly related
  - if close to 0, then they are weakly related







Covariance and Correlation, cont.

- Sample *covariance*

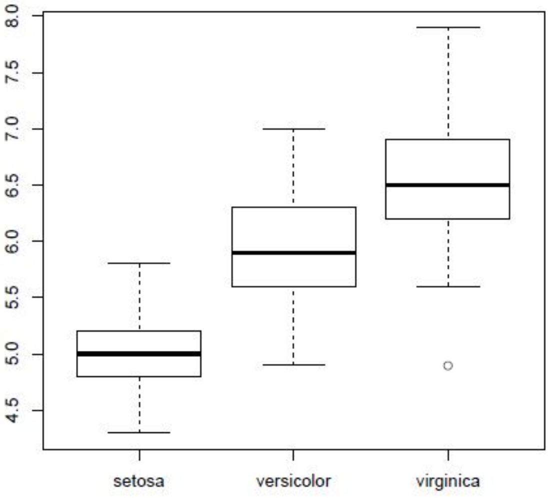
$$q_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Sample *correlation coefficient*

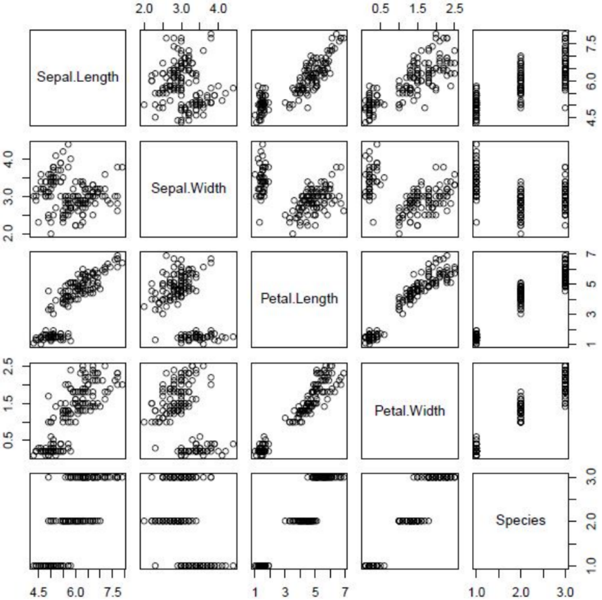
$$r_{xy} = \frac{q_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

- See Ross for details.

Exploring categorical against numerical variables



Multivariate Scatter Plots



Data Exploration Recipe

1. Find variables and decide if they are numerical or categorical.
   `str()`, `attributes()`
   - Numerical: Continuous and Discrete
   - Categorical: Ordinal and Nominal
2. Find statistics of each variable
   - Quantitative:Find `summary()`, `fivenum()`, `boxplot.stats()`
   - Qualitative: Find frequencies. `table()` or `prob.table()`
3. Perform pictorial representation of each single variable
   - Quantitative: Histograms or box plots. `hist()`, `boxplot()`
   - Qualitative: Bar or pie charts. `plot()`, `barplot()`, `pie()`
4. Be aware of outliers and robust statistics
5. Association between variables
   - `scatterplot()` to compare two numerical variables
   - Side-by-side boxplots for categorical and numerical vars
   - `pairs()` for a matrix of scatter plots of all variables
   - `cor()`, `cov()` for correlation and covariance between vars

End of Week 1