Fundamentals

Probability and Statistical Inference

Bojan Božić

TU Dublin

Winter 2020

Types of Data Analysis

- Qualitative Methods:
  - Testing theories by using language.
    - Magazine articles/intervies.
    - Conversations.
    - Documented observation.
    - Newspapers.
    - ...
- Quantitative Methods:
  - Testing theories using numbers (statistics).

What is statistical inquiry?

- The science of collecting, analyzing and drawing conclusions from data.
- The science of collecting, describing and interpreting data.

What can statistics do?

- Make data more managable.
- 6, 1, 8, 3, 5, 4, 9.
  - We can calculate an average.
  - We can state the range.
  - We can plot them on a graph.

What else can statistics do?

- Provide us with evidence from the data that allows us to draw conclusions.
  - Group of numbers #1: 6, 1, 8, 3, 5, 4, 9.
    - Average is 5.14.
  - Group of numbers #2: 8, 3, 4, 2, 7, 1, 4.
    - Average is 4.25.
- Allows us to compare groups of data.
- Allows us to do this **objectively** and **quantitatively**.

The Process (Source: Andy Field)

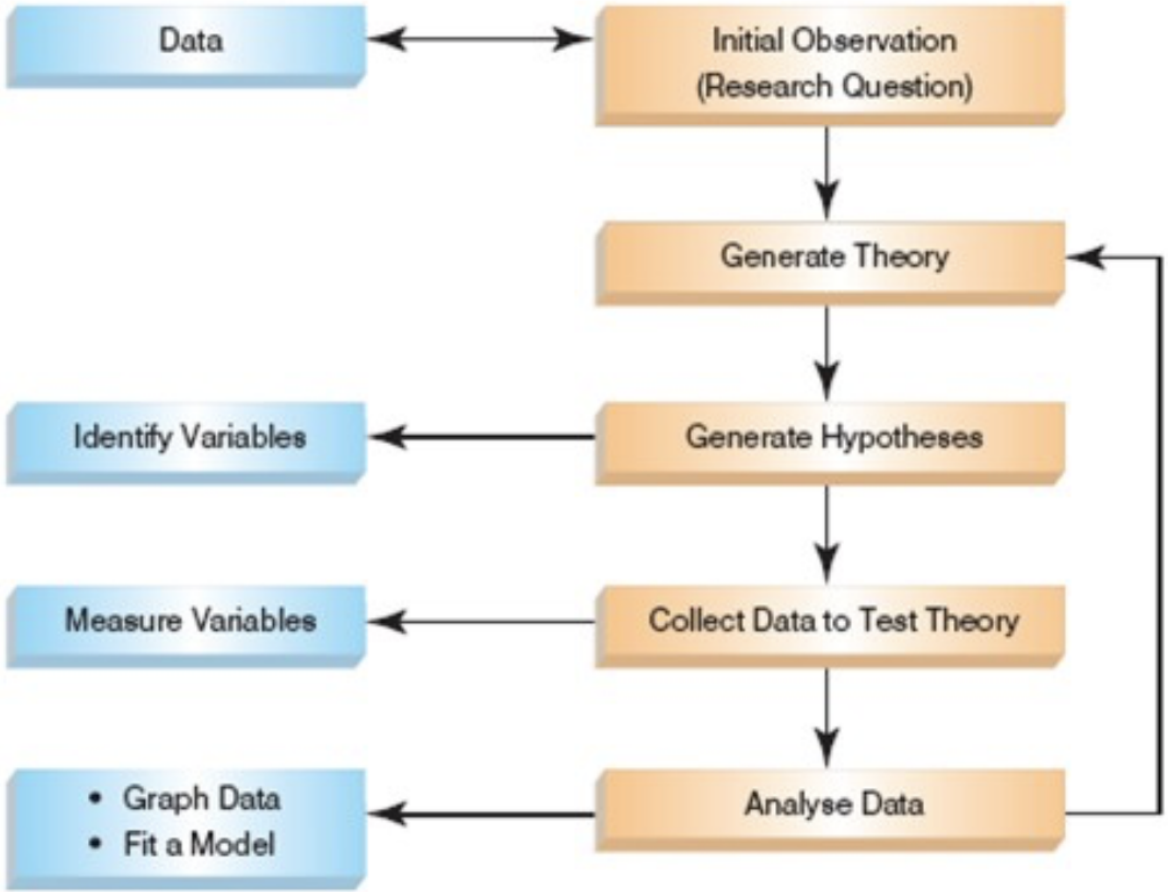


Figure: The research process.

Initial Observation

- Find something that needs explaining.
  - Observe the real world.
  - Read other research.
- Test the concept: collect data.
  - Collect data that will provide evidence you can test to see if your idea is valid.
  - To do this you need to define variables:
    - Anything that can be measured and can differ across entities or time.

The Research Process - Step 2

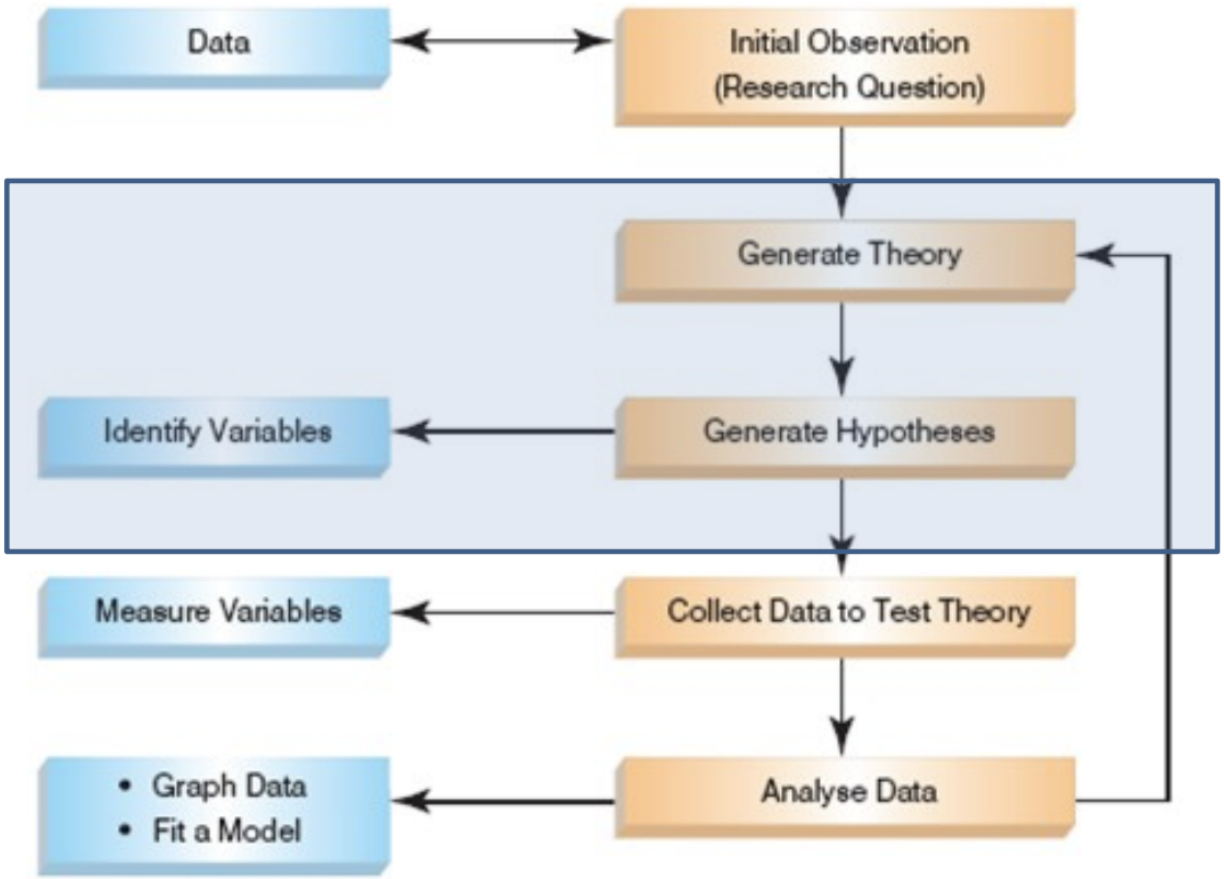
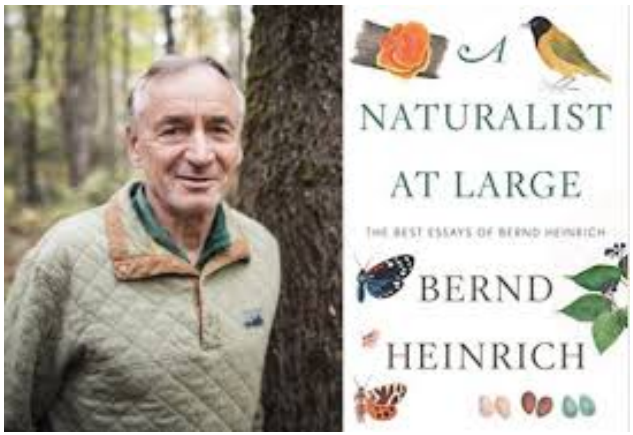


Figure: The research process (step 2).

Conceptual Framework - Why is it important?



- Bernd Heinrich
- Spent a summer conducting detailed, systematic research on ant lions.
  - Small insects that trap ants in pits they have dug.
- When he conducted his analysis he found his results were very different to other researchers. . .
  - A good thing? Was his research revolutionary?

Conceptual Framework

- Repeated experiments following summer.
- Found that he and his team had misunderstood ant lion behaviour:
  - In particular the time frame involved.
  - Missed specific behaviour that impacted the results.
- "Even carefully collected results can be misleading if the underlying context of assumptions is wrong.", Bernd Heinrich, 1984

Conceptual Framework

- System of
  - concepts,
  - assumptions,
  - expectations,
  - beliefs, and
  - theories that supports and informs your research.
- A key part of your design.
- A conception or model of what is out there that you plan to study, and of what is going on with these things and why:
  - a tentative theory of the phenomena that you are investigating.
- Influences your design and in particular the data you need to collect/use.

Generating and Testing Theories

- Theory
  - A hypothesized general principle or set of principles that explains known findings about a topic and from which new hypotheses can be generated.
    - e.g. Computer Science attracts students with strong mathematical ability.
- Hypothesis
  - A prediction from a theory.
  - E.g. the number of people applying for an MSc in Computer Science will have basic mathematical ability greater than the general level in the population.
- Falsification
  - The act of disproving a theory or hypothesis.

The Research Process - Step 3

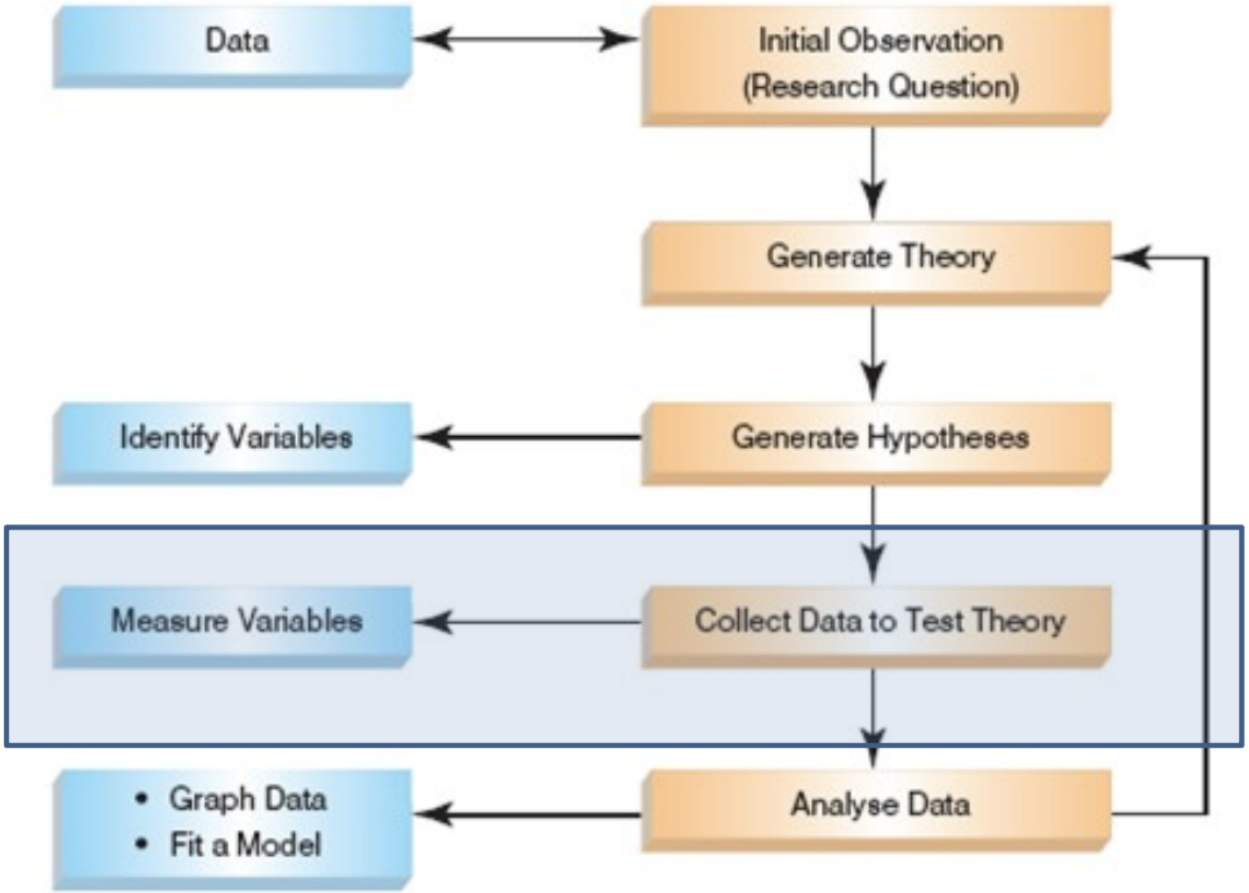


Figure: The research process (step 3).

What data do I need?

- Population
  - The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.
  - The world you are interested in and trying to build a model of.
  - Two kinds of populations:
    - finite or infinite.
- Sample
  - A portion, or part, of the population of interest.
- Statistical inquiry is usually carried out with a SAMPLE.

Population vs Sample

- Need to understand the nature of the population .
  - Need the sample to be representative.
- Factors influencing the accuracy of a sample's ability to represent a population:
  - Size
  - Randomness
- If Sample is not representative it is **biased**.

Basic Terms

- **Individuals** are the people or objects being studied
  - The set of data collected about each individual is referred to as a **case**.
- **Variable**
  - A characteristic about each individual of a population or sample.
  - Any characteristic whose value may change from one individual to another.
  - Structured to facilitate comparison between individuals.
- **Data**
  - Value of a variable for individual observations.
- **Dataset**
  - The set of values collected for the variable(s) from each of the individuals belonging to the sample/population the dataset represents.
  - Each case in a dataset has one datum or observation for each variable.

Basic Terms

- **Experiment**
  - A planned activity whose results yield a set of data.
- **Parameter**
  - A numerical value describing an aspect of an entire population.
- **Statistic**
  - A numerical value describing an aspect of the sample.

Parameter vs Statistic

- Both are descriptions of groups
- Both allow us to make statements about a group like "50% of dog owners prefer X Brand dog food."
- For example
  - You randomly poll voters in an election.
  - You find that 55% of the population plans to vote for candidate A.
  - That is a statistic.
  - Why? You only asked a sample of the population who they are voting for.
  - You calculated what the population was likely to do based on the sample.
- For small populations it is possible to measure to compute a parameter
- For larger populations you will measure to compute a statistic

Example

- A country is interested in learning about the average age of its academic staff working at third level.
- The basic terms in this situation:
- The population is all academic members working in the country and we want to determine their age.
- A sample is any subset of that population.
  - For example, we might select 50 staff members and determine their age.
  - As long as this is representative and sufficiently large to model the population.

Example

- The variable is the "age" of each staff member.
  - One datum would be the age of a specific faculty member.
- The experiment would be the method used to select the ages forming the sample and determining the actual age of each staff member in the sample.
- The parameter of interest is the "average" age of all academic staff in the country.
- The statistic is the "average" age for all academic staff in the sample.

Sample Size

- In an ideal situation, the entire population should be studied but this is almost impossible.
- Majority of studies are performed on limited subjects drawn from the concerned population known as "sample population".
- The data obtained is analysed and conclusions are drawn which are **extrapolated** to the population under study.
- Sample size - The number of cases n.
  - E.g. $n = 90$
- Too small a sample.
  - May not lead to conclusions that are valid for the population.
- Too large a sample.
  - Wasteful if a smaller sample would do.
  - May have ethical implications dependent on the type of experiment/observation being undertaken.
- How do decide on sample size is something we will talk about.

- Biased Sampling Method:
  - A sampling method that produces data which systematically differs from the population from which it is taken.
- Aim for a Simple Random Sample:
  - A sample of n measurements from a population is a subset of the population selected in such a manner that every sample of size n from the population has an equal chance of being selected.
  - Some things to note:
    - Researcher bias should not occur in the sample selection.
    - May not always reflect the diversity of the population.

Data Collection: What to measure?

- Hypothesis:
  - Consumption of Coca-Cola improves a student's ability to concentrate.
- Decide what variables you need:
  - Independent Variable
    - The proposed cause.
    - A predictor variable.
    - A manipulated variable (in experiments).
    - Coca-Cola consumption in the hypothesis above.
  - Dependent Variable
    - The proposed effect.
    - An outcome variable.
    - Measured not manipulated (in experiments).
    - A student's ability to concentrate in the hypothesis above.

Confounding Variables

- Sometimes we are interested in establishing relationships/differences and then inferring cause and effect[1].
- This can be complicated by confounding variables.
  - Variables which may or may not be directly measured and which have influence on other variables in the study.
- Two variables are confounding when one cannot be distinguished from the effects of another.
  - E.g. amount of petrol and time used to commute to work, it is likely that level of congestion will be a factor.

[1] We won't actually be able to 100% demonstrate cause and effect – we will look at this issue later.

Data Collection: Measurement Error

- Measurement error:
  - Also referred to as observational error.
  - The discrepancy between the actual value we're trying to measure, and the number we use to represent that value.
- Example:
  - Suppose you are measuring the weight of 100 athletes.
  - The scale you use is 0.1 kg out.
  - This is a systematic measurement error of 0.1 kg.
  - However, if our scale is accurate but our athletes all have different levels of hydration or are wearing different types of clothes.
    - Then we still have a measurement error – this is random.
    - There are always random errors – we need to expect them and handle them in our reporting.

Data Collection: Validity

- Whether the instrument measures what it set out to measure.
- Content validity:
  - Evidence that the content of a test corresponds to the content of the construct it was designed to cover.
- Ecological validity:
  - Evidence that the results of a study, experiment or test can be applied, and allow inferences, to real-world conditions.

Data Collection: Reliability

- Reliability
  - The ability of the measure to produce the same results under the same conditions.
- Test-Retest Reliability
  - The ability of a measure to produce consistent results when the same entities are tested at two different points in time.

Selecting variables for your study

- Looking for indicators that represent abstract concepts.
- Reliability
  - Degree to which an indicator is a consistent measuring device.
  - E.g. Is asking a student how well they did in school a reliable indicator of their ability to learn?
- Validity
  - Extent to which an indicator measures what it is intended to measure.
  - E.g. Is a student's IQ score a valid indicator of their educational achievement?

Types of Variables in Statistical Analysis

Quantitative	**Qualitative**
* Involves measurement * Data in numerical form * Objective and results in unambiguous conclusions * 5.14 vs 4.25 * 25% vs 50% * 1 hour vs 24 hours	* Describes the nature of something * Often evaluative and ambiguous * "Good" vs "Bad" * "Right" vs "Wrong" * "A lot" vs "A little" * May be numerical, e.g. Male = 1, Female = 2

Qualitative, or Attribute, or Categorical, Variable

- A variable that categorises or describes an element of a population.
- Identifies basic differentiating characteristics of the population.
- Note:
  - Arithmetic operations, such as addition and averaging, **make no sense** for data resulting from a qualitative variable.

Quantitative, or Numerical, Variable

- A variable that quantifies an element of a population.
- Observations or measurements take on numerical values.
- Note:
  - Arithmetic operations such as addition and averaging, **make sense** for data resulting from a quantitative variable.
- Discrete
  - Isolated points along a number line
  - Usually counted
  - Can only take certain values
  - E.g. rolling a dice, # students attending class
- Continuous
  - Variable that can be any value in a given range.
  - Usually measured.
  - E.g. heights of students attending class, time spent concentrating in class.

Levels of Measurement

- Nominal
  - Data can be assigned to a category.
- Ordinal
  - There is a ranking associated with the variable.
  - Data can be ordered from smallest to largest, best to worst etc.
- Interval
  - Data can be ordered but there is meaning between the values of order.
  - Allows comparison between the data values.
- Ratio
  - Data can be ordered, there are differences between the values and you can find the ratio.
  - Has an absolute 0.
  - It makes sense to say for example, one value is twice as large as the other.

Levels of Measurement

- Nominal
  - Gender: Male, Female; Country of Origin: 1=Ireland, 2=India, 3=...
- Ordinal
  - Ranking: 1^{st}, 2^{nd}, 3^{rd}, ...
  - Answer: Strongly Disagree, Disagree, Agree, Strongly Agree.
- Interval
  - Time of day on 12h clock.
- Ratio
  - Height, Weight.
  - Time on a 24h clock.
- Interval and Ration may be referred to as Scale

The Research Process - Step 4

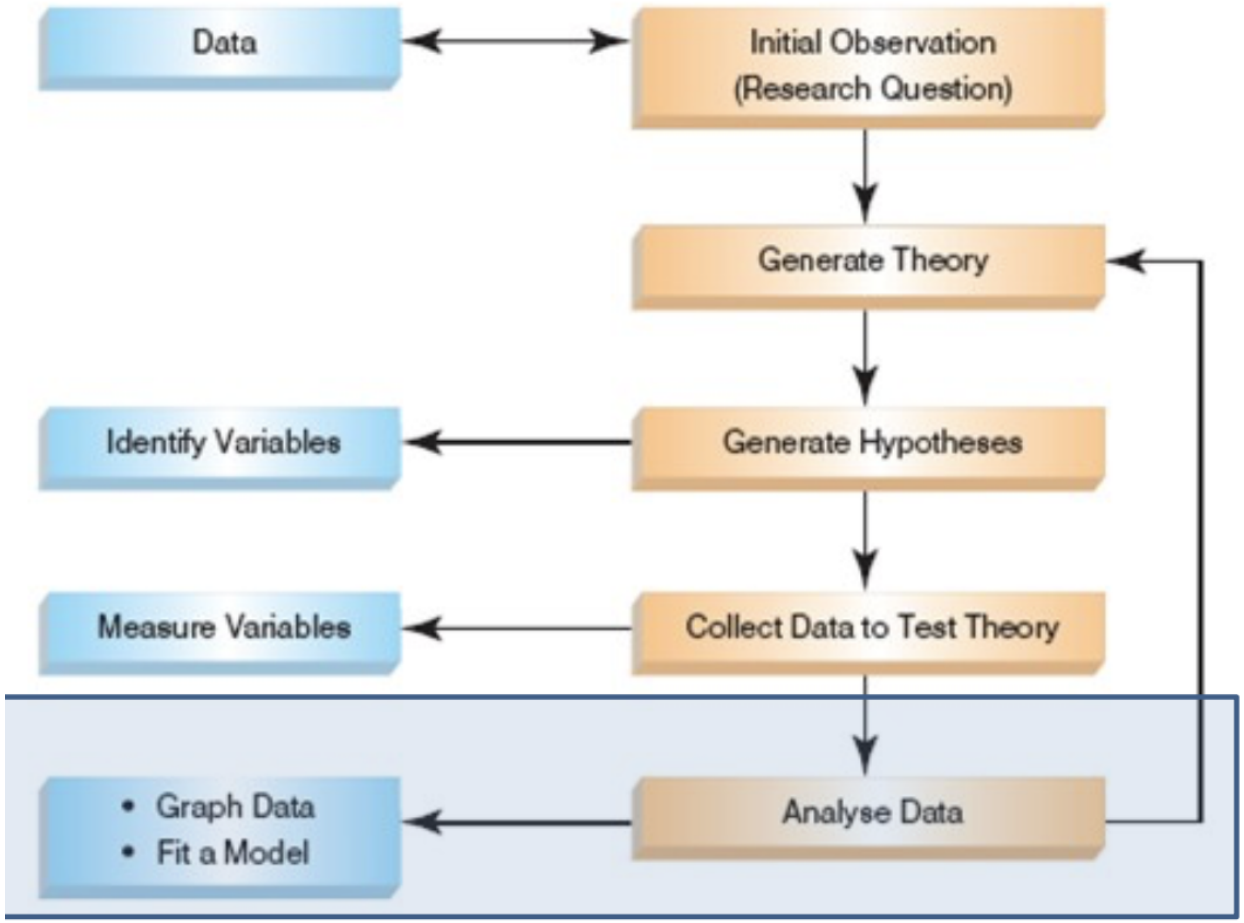


Figure: The research process (step 4).

Describing your data

- Must describe before analysing.
- Need to include sufficient detail so that your consumer can:
  - See what you are basing your analysis/conclusions on.
  - Understand everything that may constrain those analysis/conclusions.
- You are trying to present information about a large body of data so that your consumer can understand it without having to view every individual case you have collected.
  - Usually start with a picture and
  - Include some meaningful numbers to summarise and illustrate variability.

Summarising vs Analysing

- Descriptive Statistics
  - Describing the population and the sample.
  - Summarising.
- Inferential Statistics
  - Inference from sample to population.
  - Inference from statistic to parameter.

Getting Started with Statistical Study

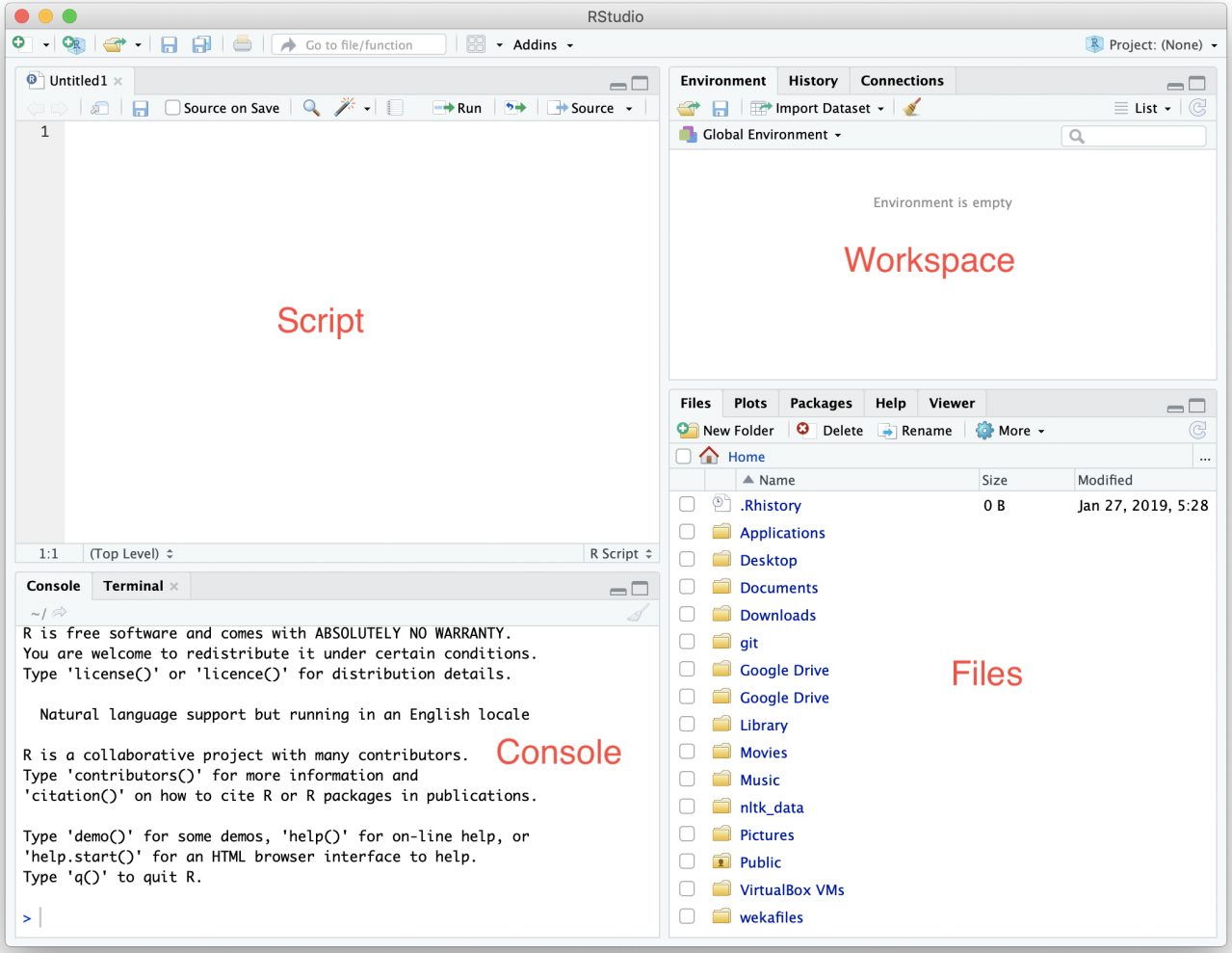
- Need to decide the nature of the study and then decide what individuals or objects are of interest.
- We must understand our population.
  - The data we need to collect.
    - Its structure and values, how it can be measured/represented.
  - Sample
    - Representativeness and size.
  - Be able to describe it in a simple way.
    - Using the appropriate type of graph for variables of interest.
- Then we will be able to analyse it appropriately.
  - Using the appropriate statistical tests to draw inference.

Time to get started ...

R Studio / R

- Launch R Studio.
- We are going to start by loading in a dataset and doing some basic descriptive statistics.
    - Create a script and run it.

R Studio Interface



R Studio Interface

- Script Window:
  - Where we open scripts with R code.
    - Can be organised into chunks.
    - We can run the commands line by line/chunk by chunk/a full file.
- Console Window:
  - Where commands get executed and results shown.
  - We can enter commands directly into the Console.

R Studio Interface

- Workspace/History Window:
  - Environment/Worksapce tab
    - A list of all objects and variables in current R environment.
    - Will contain all data files and outputs from your analysis.
    - Can save and restore this.
  - History tab
    - Contains list of every command executed in the Console.
    - Very handy backup – you can locate commands in history and sent to the Console.
- Files Window:
  1. Files - File directory structure of your current working directory,
  2. Plots - Where any plots you generate are displayed,
  3. Packages - Any packages you have loaded.
  4. Help - For every function installed in your version of R.
  5. Viewer - Used for displaying locally generated .html content.

Set Working Directory

- This is the default location where your scripts etc. will be saved and R Studio will look for datasets.
- For the moment, it is easier if you put your dataset here.
- To find it what it is currently set to (to change it):
  - In R Studio Tools $\rightarrow$ Global Options
    - In the General Menu you will see the default working directory and be able to set it.

Getting Started - Loading the Dataset

- Download data file needed:
  `https://tinyurl.com/yyw65ba4`
- Put it into your default working directory.

Getting Started - The Dataset

This is a small, simple dataset which contains the following features:

Features in the dataset

Name	**Label**	**Values**	**Type**
salary	Annual Salary		Scale
gender	Gender	1=Female, 2=Male	Nominal
rank	Rank	1=Assistant Prof. 2=Associate Prof. 3=Full Prof.	Nominal
exper	Years of experience		Scale
expcat	Years in current rank		Scale
dg	Highest degree obtained		Nominal

Note

- This tutorial is based on using the script window but you can also type these commands directly into the Console.
  - They will appear in the History Window.
  - At the end of the tutorial you can open a script and hit the To Source button on the history window to copy your commands across (select them all/chunk by chunk).

Getting Starte - Loading

In the Script Window:

- **File** $\rightarrow$ **New** $\rightarrow$ **R Script**
- In the new file type: `salary <- readRDS("salary.rds")`
- In the Script window click **Run** (top left of the screen).
- The commands will be echoed on the Console.
- In your file type in: `names(salary)`
- In the Script Window click **Run**
- The commands will be echoed in the Console and you will see the output.
- Save the file **File** $\rightarrow$ **Save As** (save to a location that you will find again).

Descriptive Statistics

```
1   # You can look at the data in a variable
2   # by entering its name at the command prompt.
3   salary$gender
4
5   # Or get a short overview using the str function.
6   str(saraly$gender)
7
8   # Or get a relevant statistical summary for a
    variable e.g. gender.
9   summary(salary$gender)
10
11  # Or for salary.
12  summary(salary$salary)
13
14  # Or get a summary of all the variables in the
    dataset.
15  summary(salary)
16
```

Frequency Tables

Can be used to illustrate frequency distribution of a categorical feature. They also tell us which values the feature takes and how often it takes them.

Gender	M	F
Frequency	14	38

Rank	Assistant	Associate	Full
Frequency	18	14	20

In R

- We need to use a package.
- We are going to use `pylr`.
- You need to install it (just once): `install.packages('pylr')`
- Then load it: `library('pylr')`
- Then create the table: `count(salary$gender)`

Contingency Table/Cross Tabulation

- Matrix used to summarise the relationship between two categorical features.
- Suppose we want to know the relationship between rank and gender - What is the probability that each gender will hold each rank?
- We create a table and then show the distributions.

```
1  tab <- table(salary$gender, salary$rank)
2  addmargins(tab)
3  tab # just display the table with frequencies.
4  prop.table(tab) # shows probability distribution.
5
```

Note: we can make these look better using some additional packages - we will look at those later.

Measure of Central Tendency

- A descriptive statistic.
- A single number to serve as a representative value around which all the numbers in the set tend to cluster.
- Need to choose the right measure for your data!

The Mode

The mode is the score that occurs most frequently in a set of data.
The value with the greatest frequency on the distribution.

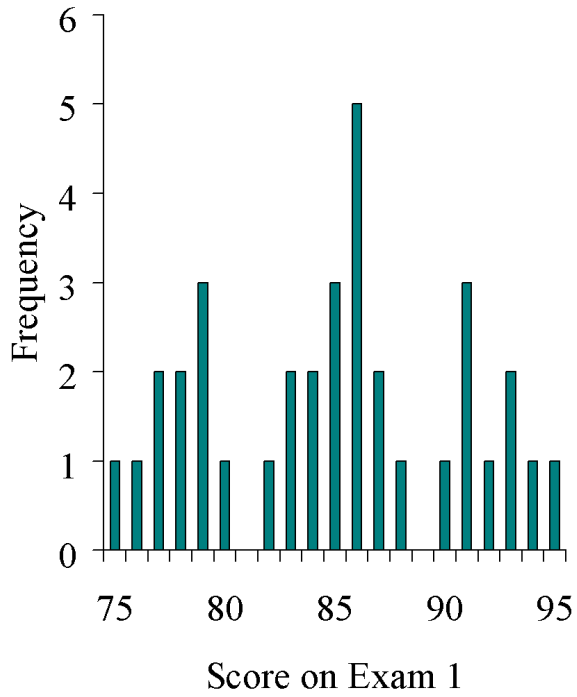


Figure: What's the mode here?

Bimodal Distributions

When a distribution has two modes, it is called *bimodal*.

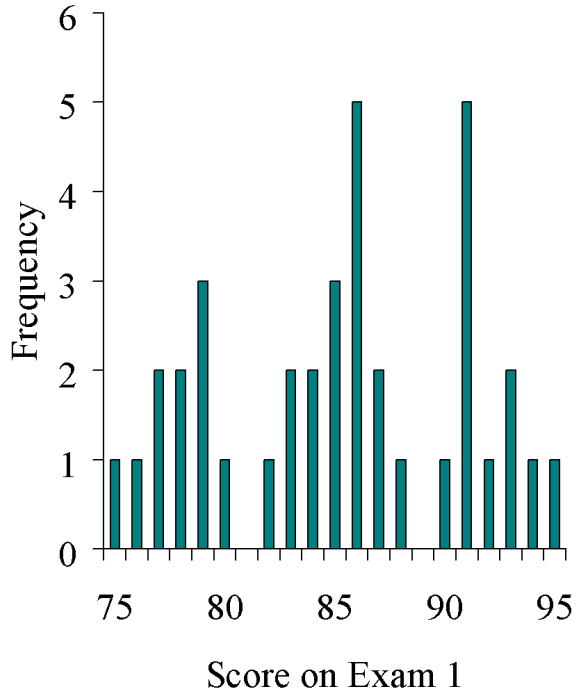


Figure: Which two values are modes here?

Multimodal Distributions

If a distribution has more than 2 modes, it is called *multimodal*.

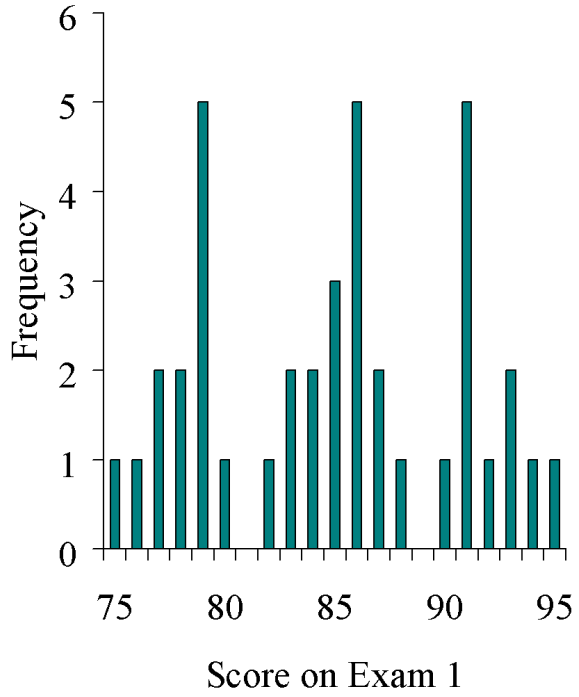


Figure: Multimodal distributions have more than two modes.

When to use the mode?

- The mode is not a very useful measure for central tendency.
- It is insensitive to large changes in the dataset.
  - That is, two datasets that are very different from each other can have the same mode.

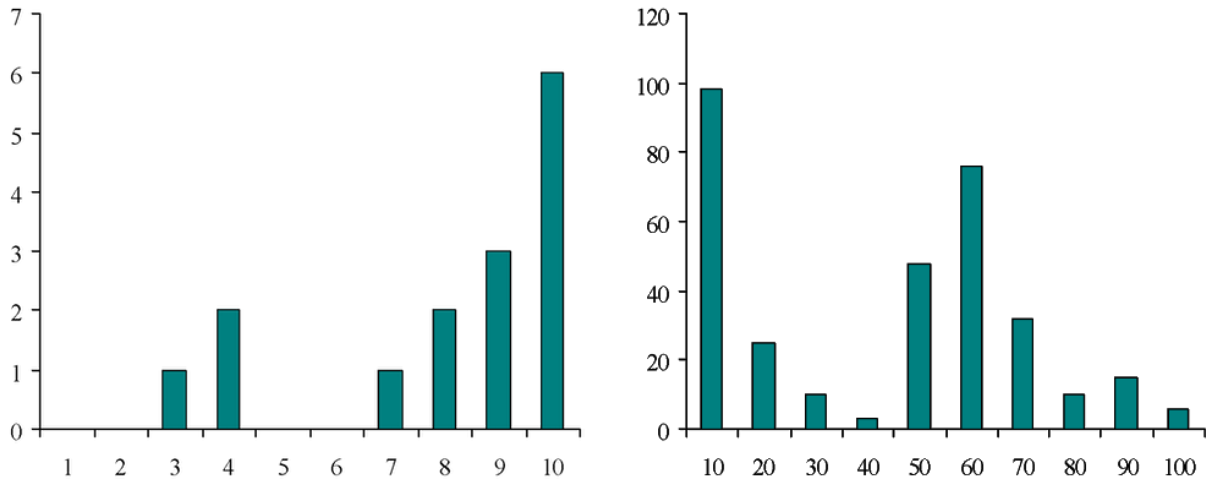


Figure: Different datasets, same mode.

When to use the mode?

- The mode is primarily used with nominal variables.
- Example:
  - 3, 7, 5, 13, 20, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29.
  - If we order the list:
  - 3, 5, 7, 12, 13, 14, 20, **23, 23, 23, 23,** 29, 39, 40, 56.
  - In this case, the mode is 23.

Back to Salary in R Studio

R doesn't have a built in function to compute the mode but there is a simple function which you can write and include in your code that will do it for you:

```
1     getmode <- function(v)
2     {
3         uniqv <- unique(v)
4         uniqv[which.max(tabulate(match(v,uniqv)))]
5     }
6
```

Using the function to get the mode of salary:

```
1     getmode(salary$salary)
2
```

The Median

- Just another name for the 50th percentile.
  - It is the score in the middle; half of the scores are larger than the median and half of the scores are smaller.
  - The middle score of a sequence of all the scores in a distribution ordered from lowest to highest.
- Sort the data from highest to lowest.
- Find the score in the middle.
  - `middle = (N+1)/2`
  - If N, the number of scores, is even, the median is the average of the middle two scores.

Median Example

- What is the median of the following scores:
  10, 8, 14, 15, 7, 3, 3, 8, 12, 10, 9.
- Sort the scores:
  15, 14, 12, 10, 10, 9, 8, 8, 7, 3, 3.
- Determine the middle score:
  `middle = (N+1)/2 = (11+1)/2 = 6`.
- Middle score = median = 9.

Median Example

- What is the median of the following scores:
  24, 18, 19, 42, 16, 12.
- Sort the scores:
  42, 24, 19, 18, 16, 12.
- Determine the middle score:
  `middle = (N+1)/2 = (6+1)/2 = 3.5`.
- Median is average of 3^{rd} and 4^{th} scores: $(19+18)/2 = 18.5$.

The Mean

The arithmetic average of a group of scores; the sum of the scores divided by the number of scores.

Calculating the Mean

- Calculate the mean of the following data:
  1, 5, 4, 3, 2.
- Sum the scores ($\sum X$):
  $1+5+4+3+2 = 15$.
- Divide the sum ($\sum X = 15$) by the number of scores ($N = 5$):
  $15/5 = 3$.
- Mean = X = 3.
- The mean is sensitive to extreme values.

Mean, Mode, Median

- Does it make a difference?
- UK Salary
  - Mean annual earnings £24,000 - £463 per week.
  - Median earnings £377 per week.
  - Mode earnings £270 per week (approx).

Measures of Central Tendency

Need to be considered in relation to the variability within the dataset.

Measures of Dispersion

- Which of the distributions of scores has the larger dispersion?
- The distribution to the left has more dispersion because the scores are more spread out.
- That is, they are less similar to each other.

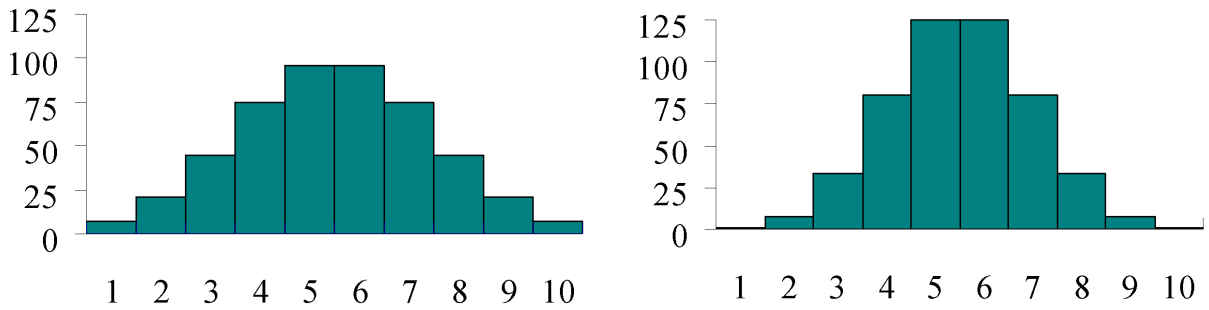


Figure: Comparison of measures of dispersion.

Measures of Dispersion

Descriptive statistics that describe how similar a set of scores is to another.

- The more similar the scores are to each other, the lower the measure of dispersion will be.
- The less similar the scores are to each other, the higher the measure of dispersion will be.
- In general, the more spread out a distribution is, the larger the measure of dispersion will be.

Measures of Dispersion

There are three main measures of dispersion:

- The range.
- The semi-interquartille range (SIR).
- Variance / standard deviation.

The Range

- The range is defined as the difference between the largest score in the set of data and the smallest score in the set of data.
  - Depends only on extreme values and provides no information about how remaining data is distributed.
- What is the range of the following data:
  4, 8, 1, 6, 6, 2, 9, 3, 6, 9.
- The largest score (XL) is 9; the smallest score (XS) is 1; the range is XL - XS $=$ 9 - 1 $=$ 8.

When to use the range?

- The range is used mainly for ordinal data.
- The range is rarely used for scientific work as it's fairly insensitive.
  - It depends on only two scores in the set of data, XL and XS.
  - Two very different sets of data can have the same range.

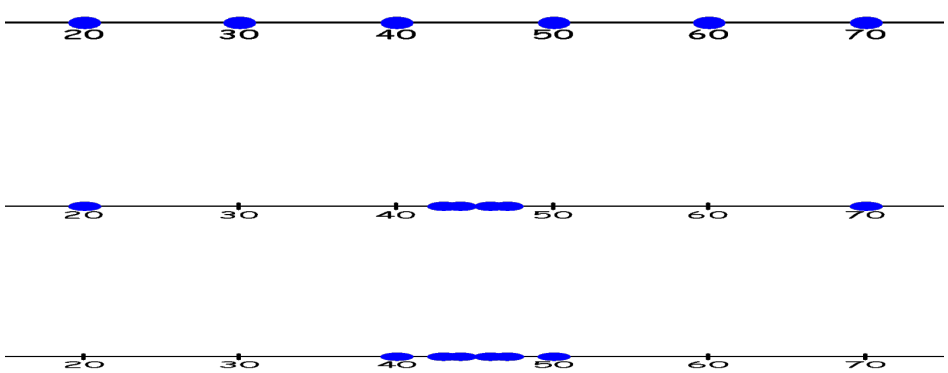


Figure: Range and data.

What is Variance?

- Concerned with deviations from the mean (X-μ).
- Actually the average of the **squared** differences from the mean.
- First subtract the mean from each of the scores.
  - This difference is called a *deviate* or a *deviation* score.
  - The deviate tells us how far a given score is from the typical, or average, score.
  - Thus, the deviate is a measure of dispersion *for a given score*.
- Then *square* the result.
  - Why?
  - If we just added up the differences from the mean the negatives would cancel the positives.
  - If we used absolute values we wouldn't get an accurate measure of spread.
  - Squaring is the best option.

Variance

- *Variance* is defined as the average of deviates of the mean squared: $\sigma^2 = \frac{\sum(X-\mu)^2}{N}$
- N here is the degrees of freedom = the number of independent pieces of information on which the estimate is based.

Standard Deviation

- Concept was first introduced by Karl Pearson in 1893.
- The standard deviation is the most useful and the most popular measure of dispersion.
- Use Greek symbol sigma σ.
- We calculate the variance.
- The square root of this sum is known as standard deviation.
- Standard deviation $=$ the **square root** of the **Variance**.

Standard Deviation

- The larger the value the more spread out around the mean the data will be, smaller means less spread.
- The Empirical Rule:
  - The 68-95-99.7 Rule.
  - In the normal distribution with mean μ and standard deviation σ:
    - 68% of the observations fall within σ of the mean μ.
    - 95% of the observations fall within 2σ of the mean μ.
    - 99.7% of the observations fall within 3σ of the mean μ.
  - Allows us to see how spread out individual cases are from the mean.

Standard Deviation

This is helpful in comparing samples:

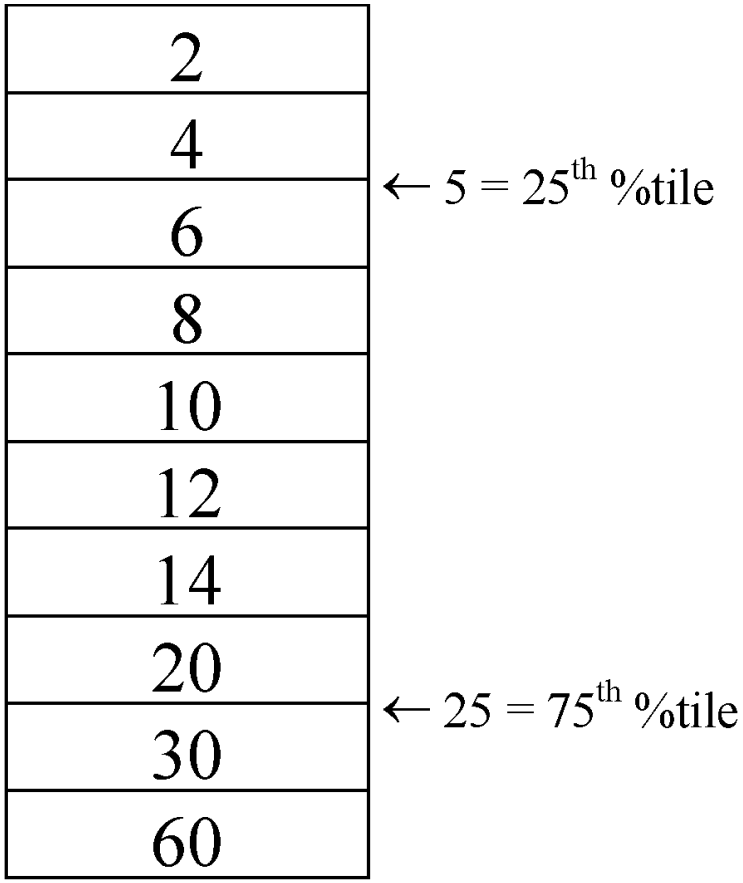
- Cork IT and TU Dublin have the same mean test score for Probability and Statistics with the same number of students in each data set.
- Does this mean that students performed equally well?
- If the mean for TU Dublin is 60 and the standard deviation is 1.6 then
  - 68% of the values in the dataset will lie between 58.4 and 61.6
    - MEAN-1SD (60-1.6=58.4) and MEAN+1SD (60+1.6=61.6)
  - 99% of the values will lie between 55.2 and 64.8
    - MEAN-3SD (60-4.8=55.2) and MEAN+3SD (60+4.8=64.8)
- If the mean for CIT is 60 and the standard deviation is 4.3
  - 68% of the values in the dataset will lie between 55.7 and 64.3
    - MEAN-1SD (60-4.3=55.7) and MEAN+1SD (60+4.3=64.3)
  - 99% of the values will lie between 47.1 and 72.9
    - MEAN-3SD (60-12.9=47.1) and MEAN+3SD (60+12.9=72.9)

The Interquartile Range

- The Interquartile range (IQR) is defined as the difference of the first and third quartiles divided by two.
  - The first quartile is the 25th percentile.
  - The third quartile is the 75th percentile.
- IQR = (Q3 - Q1) / 2

Example

- What is the IQR for the data to the right?
- 25% of the scores are below 5.
  - 5 is the first quartile.
- 25% of the scores are above 25.
  - 25 is the third quartile.
- IQR= (Q3 - Q1) / 2 = (25 - 5) / 2 = 10.



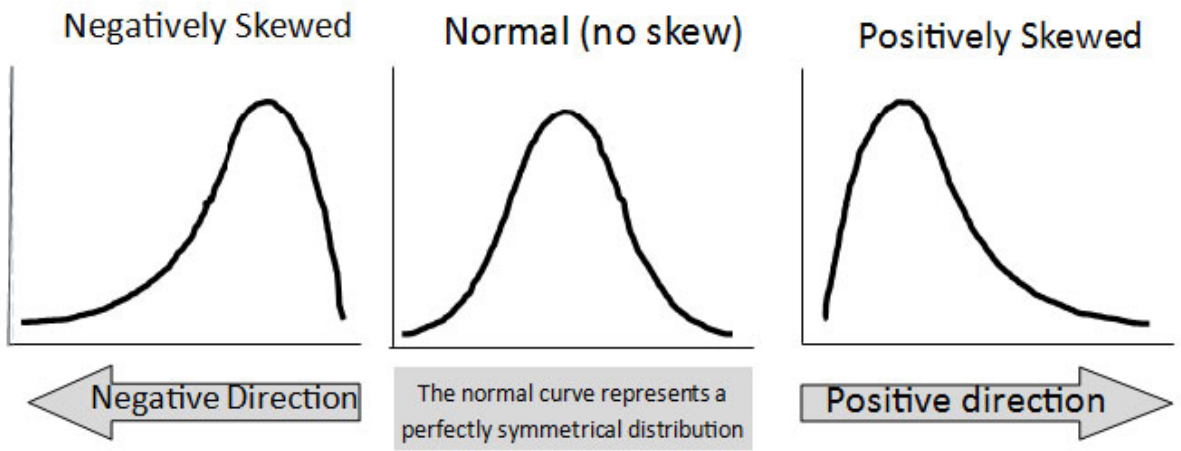
Which measures should we use?

It depends on:

1. The expectations of your area.
2. On the size of your sample.
3. On the distribution of your data
   - Need always to look at this and then make a decision whether 1 and 2 allow you to consider your data to fit the normal distribution.

Measure of Skew

- Skew is a measure of symmetry in the distribution of scores.
- If the curve has one tail that is longer than the other, it is *skewed*.
- If the longer tail is on the left, it is *negatively skewed* (too many scores toward the *negative* end).
- If the longer tail is on the right, it is *positively skewed*.



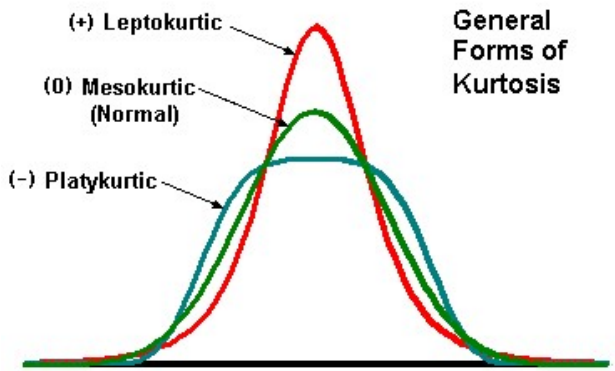
Relations Between the Measures of Central Tendency

- In symmetrical distributions, the median and mean are equal.
- For normal distributions, mean = median = mode.
- In positively skewed distributions, the mean is greater than the median.
- In negatively skewed distributions, the mean is smaller than the median.

Kurtosis

Kurtosis measures whether the scores are spread out more or less than they would be in a normal (Gaussian) distribution.

- When the distribution is normally distributed, its kurtosis equals 3 and it is said to be *mesokurtic*.
- When the distribution is less spread out than normal, its kurtosis is greater than 3 and it is said to be *leptokurtic*.
- When the distribution is more spread out than normal, its kurtosis is less than 3 and it is said to be *platykurtic*.



When To Use the Median & IQR

- The median is often used when the distribution of scores is either positively or negatively skewed.
  - The few really large scores (positively skewed) or really small scores (negatively skewed) will not overly influence the median.
- The IQR is often used with skewed data as it is insensitive to the extreme scores.

When To Use the Mean and Standard Deviation

- You should use the mean when.
  - The data is interval or ratio scaled.
    - Many people will use the mean with ordinally scaled data too.
  - and the data are not skewed.
- The mean is preferred because it is sensitive to every score.
  - If you change one score in the data set, the mean will change.
- If you use mean you also use standard deviation.

Back to Salary in R Studio

- Measures of Central Tendency: `mean(salary$salary)`
- You can assign the outcome to a variable: `meansal <- mean(salary$salary)`
- and then display it on screen: `meansal`
- Or use the print function to make it look the way you want: `print(meansal, digits=1)`
- Median: `median(salary$salary)`

Measures of Dispersion

```
# Range
range(salary$salary)

# Quantiles
quantile(salary$salary)
# to get 1st quantil
x=quantile(salary$salary); x[1]
# Interquartile Range
IQR(salary$salary)

# Variance
var(salary$salary)
# Standard deviation
sd(salary$salary)
# Rounded
round(sd(salary$salary),2)
# rounded to 2 decimal places

```

Some Key Concepts

- Sample: Size, representativeness, extremes;
- Differing types of variable: nominal/categorical, ordinal, interval and ratio;
- Measures of central tendency: the mean, median and mode;
- Measures of dispersion: range, inter-quartile range, variance and standard deviation;
- Shape: normal, skew, kurtosis.