Data Science

- Turning **data** into something meaningful
- **Science** of uncertainty
- Quintessential **interdisciplinary** science

Data Science Skillset

- Statistics, mathematics and IT skills (e.g. programming)

Data Science Skillset

- **Statistics**, mathematics and IT skills (e.g. **programming**)

Data Science Skillset

- **Statistics**, mathematics and IT skills (e.g. **programming**)
- Logical thinker
- Problem solver
- Good communicator

What does the term,

“statistics ”,

mean to you ?

What is / are Statistics?

What is / are Statistics?

A statistic:

Science of statistics:

What is / are Statistics?

A statistic[*]: any quantity computed from sample data

Science of statistics:
collecting, classifying, summarizing, organizing, analyzing, estimation and interpretation of information

[*] Terminology also used for function to calculate the summary quantity

Role of Statistics

Field of statistics deals with the collection, presentation, analysis, and use of data to:

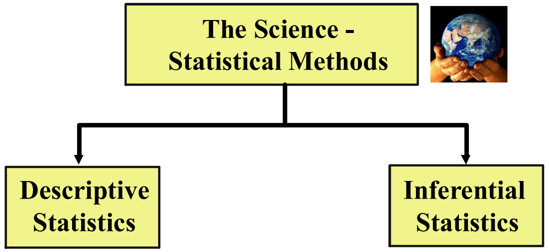
- make decisions
- solve problems
- design products and processes

It is the science of uncertainty

Role of Probability

•Probability provides the **framework** for the study and application of statistics

Statistical Methods



The Science - Statistical Methods

Descriptive Statistics

Inferential Statistics

https://www.nuigalway.ie/adult-learning/about-us/didyouknow/

https://visual.ly/community/infographic/animals/shark-attack

Descriptive Statistics: *Science of summarizing data, numerically and graphically…*

Analysis methods applicable depends on the variable being measured and the research questions which you are trying to answer …

<span style="color:orange">**Inferential Statistics:**</span> *science of using the **information in your sample** to say (i.e. to "**infer**") something **about the population** of interest*

Thinking Challenge

Suppose the student newspaper is interested in what proportion of NUI Galway students pay rent

and

the average amount of rent paid

How would you find out?

Breakdown the question...

What is the individual / experimental unit?

What is the population of interest?

What are the variables of interest?

What types are these variables?

What are the parameters of interest?

How would you collect the data?

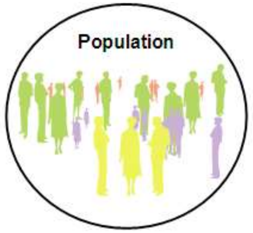
What are the observations for the variables?

How would you summarise these observations?

Some important terms:

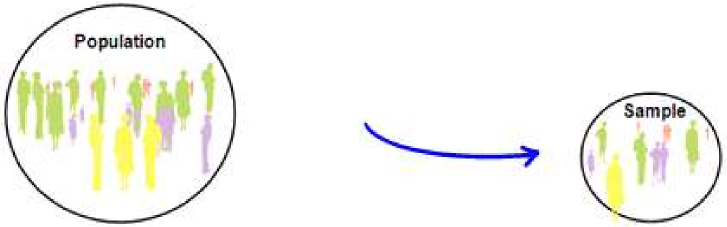
An ***experimental unit / individual***
is a single object upon which we collect data,
e.g. person,
thing,
transaction,
event.

Some important terms:



A ***population***
is a collection of
experimental units/individuals
that we are interested in studying.
e.g. people,
things,
transactions,
events

Some important terms:



Some important terms:

A **sample**
is a subset of experimental units /
individuals from the population.
e.g. people,
things,
transactions,
events

Some important terms:

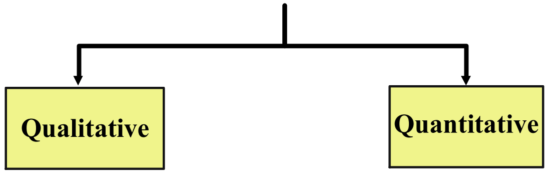
A **variable** is a characteristic or property of an individual experimental unit.

examples:

height
grade score
account balance
gender (m/f/non-binary),
letter grade (A, B, C, etc.),
Likert scale (agree, neutral, disagree, etc.)

Types of variable:

A **variable** is a characteristic or property of an individual experimental unit


Qualitative

Quantitative

May be measured, or more generally "observed", on each individual

Qualitative Data:

Classified into categories, can be **ordered**:

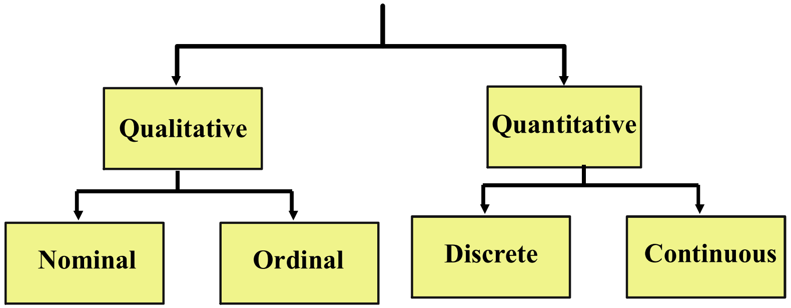
- Grade achieved in ST2001

or **unordered**:

- Gender of each employee at a company
- Method of payment (cash, cheque, credit card)

Types of variable:

A **variable** is a characteristic or property of an individual experimental unit.


Qualitative

Quantitative

Nominal

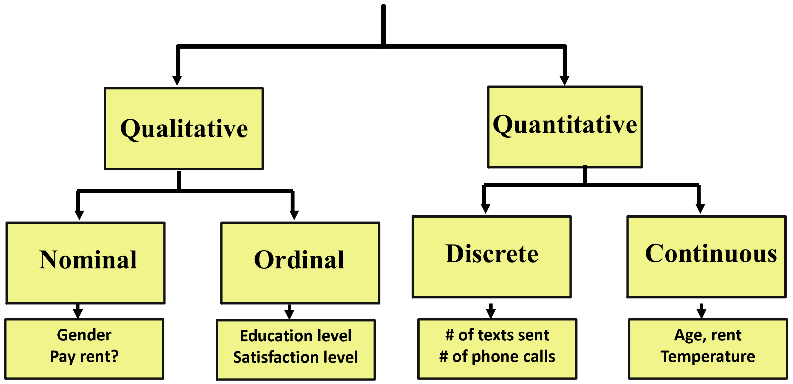
Ordinal

Discrete

Continuous

Types of variable:

A **variable** is a characteristic or property of an individual experimental unit.


Qualitative

Quantitative

Nominal

Ordinal

Discrete

Continuous

Gender
Pay rent?

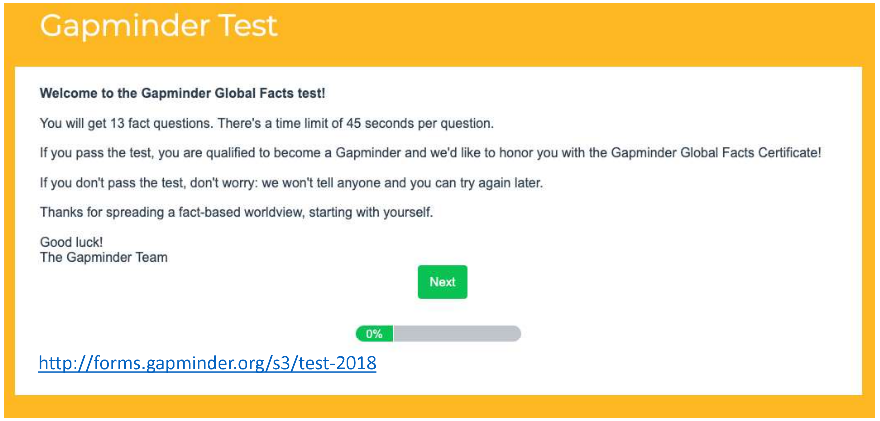
Education level
Satisfaction level

# of texts sent
of phone calls

Age, rent
Temperature

Gapminder Data: https://www.gapminder.org/

The Gapminder Foundation is a Swedish NGO which promotes sustainable global development by increased use and understanding of statistics about social, economic and environmental development



Gapminder Test

Welcome to the Gapminder Global Facts test!

You will get 13 fact questions. There's a time limit of 45 seconds per question.

If you pass the test, you are qualified to become a Gapminder and we'd like to honor you with the Gapminder Global Facts Certificate!

If you don't pass the test, don't worry: we won't tell anyone and you can try again later.

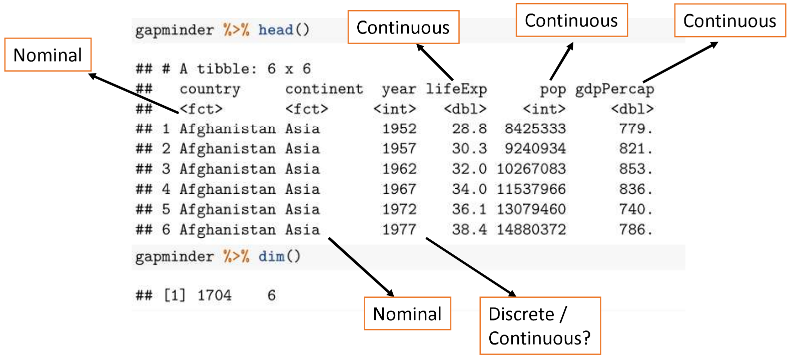
Thanks for spreading a fact-based worldview, starting with yourself.

Good luck!
The Gapminder Team

http://forms.gapminder.org/s3/test-2018

- What is the *typical observation*?

Gapminder Data



```
gapminder %>% head()

## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
gapminder %>% dim()

## [1] 1704    6
```

- What is the *typical observation*?
- Is there much *variation/spread* between individuals in the dataset?

- What is the *typical observation*?
- Is there much *variation/spread* between individuals in the dataset?
- How are the observations distributed over all individuals in the group – i.e. what is the shape or *distribution*?

- What is the *typical observation*?
- Is there much *variation/spread* between individuals in the dataset?
- How are the observations distributed over all individuals in the group – i.e. what is the shape or *distribution*?
- Are there any values lying outside of the range where the majority of the dataset values lie – *outliers*?

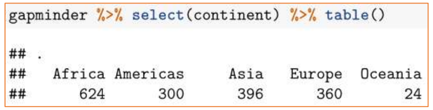
- What is the *typical observation*?
- Is there much *variation/spread* between individuals in the dataset?
- How are the observations distributed over all individuals in the group – i.e. what is the shape or *distribution*?
- Are there any values lying outside of the range where the majority of the dataset values lie – *outliers*?

Summarising data (variables) can be done **numerically**, with appropriate summaries, or **graphically**, with appropriate plots

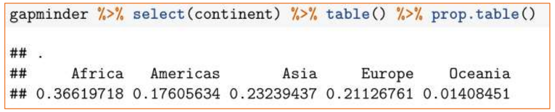
Summarising Categorical Data

- **Numerical Summary:** frequency count and percentage

Continent	Frequency	Proportion
Africa	624	0.36619718
Americas	300	0.17605634
Asia	396	0.23239437
Europe	360	0.21126761
Oceania	24	0.01408451



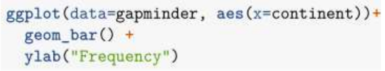
```
gapminder %>% select(continent) %>% table()

## .
##   Africa Americas     Asia   Europe  Oceania
##      624      300      396      360       24
```



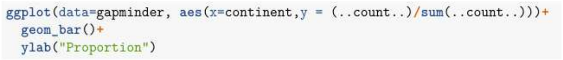
```
gapminder %>% select(continent) %>% table() %>% prop.table()

## .
##     Africa   Americas       Asia     Europe    Oceania
## 0.36619718 0.17605634 0.23239437 0.21126761 0.01408451
```

Summarising Categorical Data

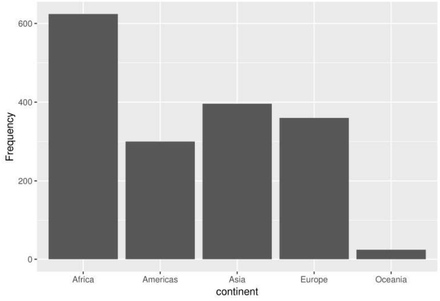
- Graphical summary: bar chart, pie chart

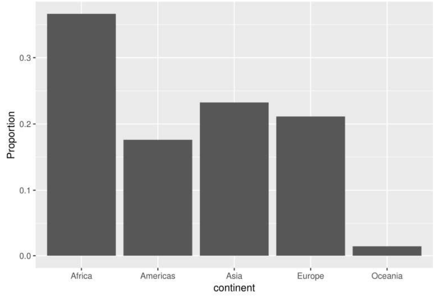


```
ggplot(data=gapminder, aes(x=continent))+
  geom_bar() +
  ylab("Frequency")
```



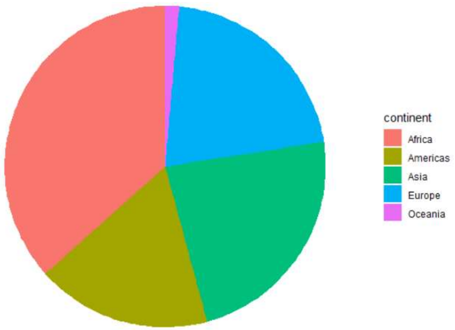
```
ggplot(data=gapminder, aes(x=continent,y = (..count..)/sum(..count..)))+
  geom_bar()+
  ylab("Proportion")
```





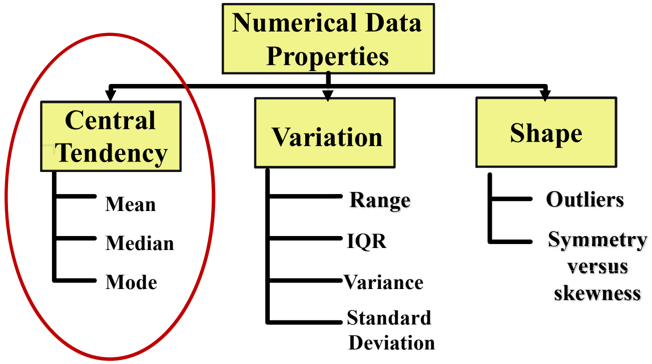
Summarising Categorical Data

- Graphical summary: bar chart, pie chart



Advice: don’t use pie charts
People find determining angles very difficult
Easier to understand lengths/heights

Summarising Continuous Data



Numerical summary of typical value:

Definition

Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The **sample mean**, denoted by $\bar{x}$, is

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Sensitive to extreme values

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the sample median is

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$$

NOT Sensitive to extreme values

Mode is the most frequent observation in a dataset.

Example

Data: breaking strength of wire in kilograms
220 214 222 218 223 210 223 210 227 225 212

- Find the median:
  - Order the data from lowest to highest

210 210 212 214 218 220 222 223 223 225 227

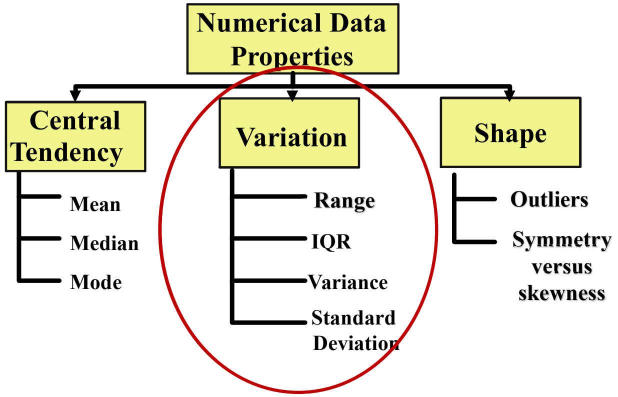
Median

- Find the Mean:

$$Mean = \frac{220+214+\ldots+222}{11} = 218.5455$$

- Mode is 210 and 223, as both have been repeated twice

Summarising Continuous Data



Numerical Summary of Spread

- Range = ***maximum - minimum***

Examples:

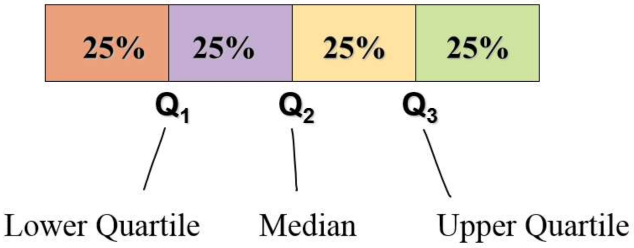
- 1, 2, 5, 8, 10 gives range of 10 – 1 = 9
- 1, 5, 5, 5, 10 also gives range of 9

- Clearly the range is poor measure of spread
- Also badly affected by outliers

Numerical Summary of Spread

- Interquartile range (IQR = Q_3 - Q_1)
- Middle 50% range of data, so is robust to outliers

Split ordered data into 4 quarters



Tukey's Method for IQR (lots of others)

Data: breaking strength of wire in kilograms
220 214 222 218 223 210 223 210 227 225 212

Put data in ascending order:

210 210 212 214 218 220 222 223 223 225 227

Q_1 = 213 Median Q_3 = 223

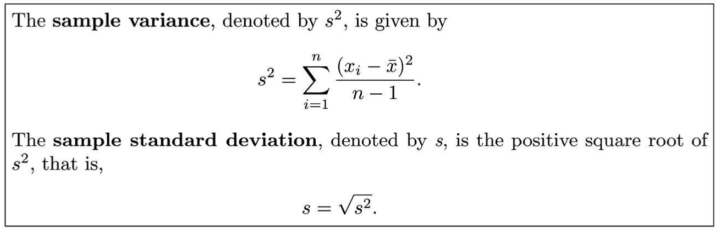
Lower (Upper) quartile is median of lower (upper) 50% of data including the median

$$IQR = Q_3 - Q_1 = 223 - 213 = 10$$

Numerical Summary of Spread

- Common measure of spread is the standard deviation, which takes into account how far *each* data value is from the mean
- A deviation is the distance of a datapoint from the mean
- Since the sum of all the deviations would be zero, we square each deviation and find an average (of sorts) of them (called the **variance**)
- We the square-root this average squared deviation… **Why?**

Definition



The **sample variance**, denoted by s^2, is given by

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}.$$

The **sample standard deviation**, denoted by s, is the positive square root of s^2, that is,

$$s = \sqrt{s^2}.$$

Sample Standard Deviation

- In same units as original variable
  - So preferable to sample variance, which is in squared units
- But… it is sensitive to outliers

Example

Data: breaking strength of wire in kilograms
220 214 222 218 223 210 223 210 227 225 212

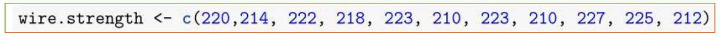
- **Find the sample variance**
- **Find the sample standard deviation**

$$\bar{x} = 218.5455$$

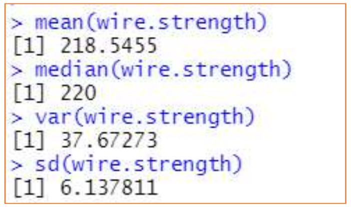
$$Sample\ Variance = s^2 = \frac{(220-218.5455)^2 + (214-218.5455)^2 + \cdots + (222-218.5455)^2}{11-1} = 37.67273$$

$$Sample\ Standard\ deviation = s = \sqrt{Sample\ Variance} = \sqrt{37.67273} = 6.1378$$

Numerical Summary in R: Vector

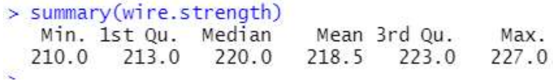


```
wire.strength <- c(220,214, 222, 218, 223, 210, 223, 210, 227, 225, 212)
```



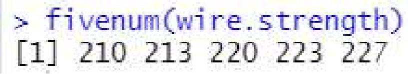
```
> mean(wire.strength)
[1] 218.5455
> median(wire.strength)
[1] 220
> var(wire.strength)
[1] 37.67273
> sd(wire.strength)
[1] 6.137811
```

summary() function uses a different formula for quartiles



```
> summary(wire.strength)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  210.0   213.0   220.0   218.5   223.0   227.0
```

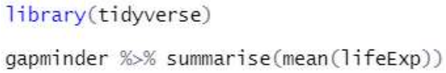
fivenum() function uses Tukey's method for Q_1 and Q_3, called the five number summary



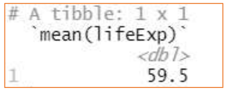
```
> fivenum(wire.strength)
[1] 210 213 220 223 227
```

Numerical Summary in R:

Calculate the **mean** of life expectancy for gapminder data:

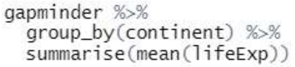


```
library(tidyverse)

gapminder %>% summarise(mean(lifeExp))
```

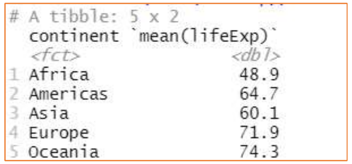


```
# A tibble: 1 x 1
  `mean(lifeExp)`
            <dbl>
1            59.5
```

Calculate the **mean** of life expectancy for different continents:

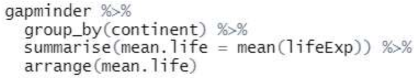


```
gapminder %>%
  group_by(continent) %>%
  summarise(mean(lifeExp))
```

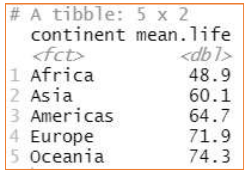


```
# A tibble: 5 x 2
  continent `mean(lifeExp)`
  <fct>               <dbl>
1 Africa               48.9
2 Americas             64.7
3 Asia                 60.1
4 Europe               71.9
5 Oceania              74.3
```

arrange

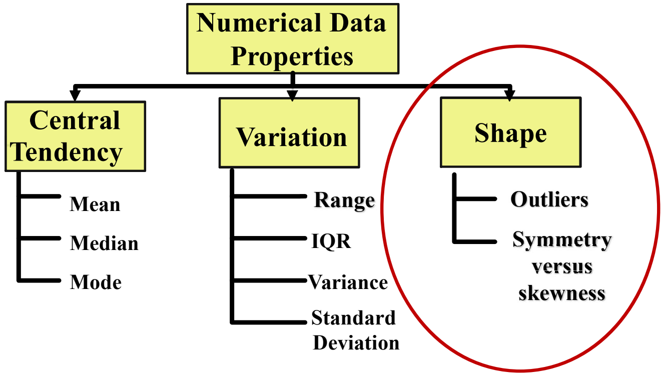


```
gapminder %>%
  group_by(continent) %>%
  summarise(mean.life = mean(lifeExp)) %>%
  arrange(mean.life)
```



```
# A tibble: 5 x 2
  continent mean.life
  <fct>         <dbl>
1 Africa         48.9
2 Asia           60.1
3 Americas       64.7
4 Europe         71.9
5 Oceania        74.3
```

Summarising Continuous Data



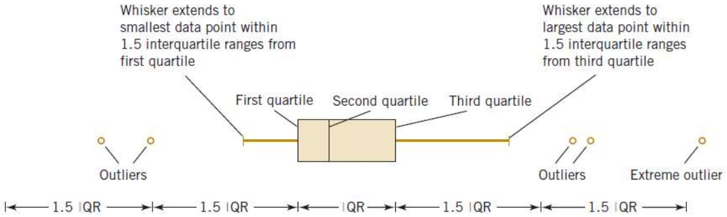
Summarising Continuous Data: Shape

- Graphical summary: boxplot, histogram

Boxplot

- A boxplot is a graphical display showing center, spread, shape, and outliers.
- It displays the 5-number summary:

min, Q_1, *median*, Q_3, and *max*

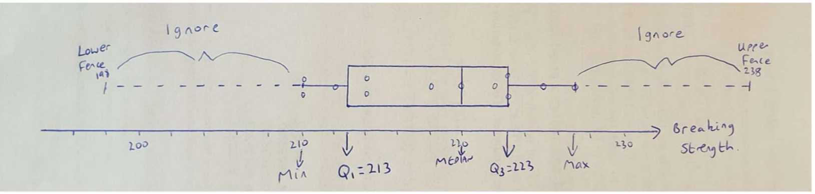

Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

Boxplot of Breaking Length

Data: breaking strength of wire in kilograms
220 214 222 218 223 210 223 210 227 225 212

```
Variable           Minimum      Q1  Median      Q3  Maximum
Breaking Length     210.00  213.00  220.00  223.00   227.00
```

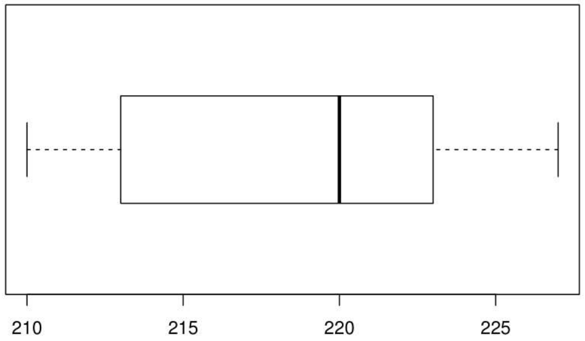
Upper fence: Q_3 + 1.5 IQR = 223 + 1.5 × 10 = 238
Lower fence: Q_1 – 1.5 IQR = 213 – 1.5 × 10 = 198



Think about a garden "fence" and closest ball is within your gardenn!

Graphical Summary in R: `boxplot()`

```
x = c(220, 214, 222, 218, 223, 210, 223, 210, 227, 225, 212)
boxplot(wire.strength, horizontal=TRUE)
```



- Note: `boxplot()` function in `R` gives exactly same result
- Other functions / software may use different method to calculate the quartiles (and/or fences)
- Usually these differences are minor so can be ignored

Histograms

✓ Useful to show general shape, location and spread of data values – representation by ***area***

Construction

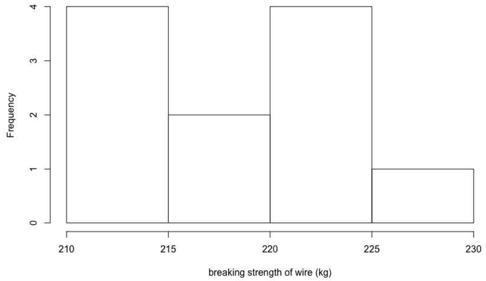
- Determine range of data – *minimum*, *maximum*
- Split into convenient intervals (or bins)
- Usually use 5 to 15 intervals
- Count number of observations in each interval - *frequency*

Histogram of Breaking Length

Data: breaking strength of wire in kilograms

220 214 222 218 223 210 223 210 227 225 212

- **Find the minimum and maximum**
- **Make classes of width 5 starting from minimum**
- **Count the frequency**
- **Plot the histogram!**



Shape of the data

When talking about the shape of the data, make sure to address the following three questions:

1. Does the histogram have a single, central hump or several well separated bumps?
2. Is the histogram or boxplot symmetric? Or more spread out in one direction, i.e. skewed
3. Any unusual features? e.g. outliers, spikes

Features to look for

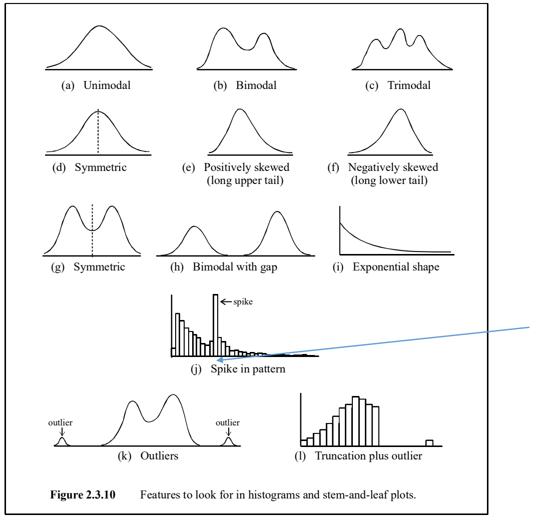


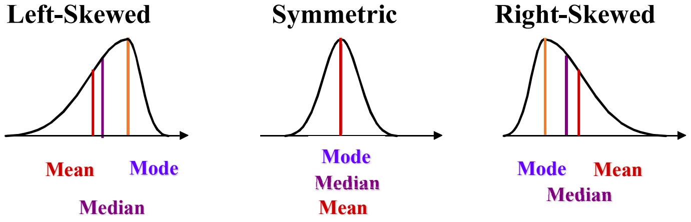
Figure 2.3.10 Features to look for in histograms and stem-and-leaf plots.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber © John Wiley & Sons 2000

e.g. minimum value
for free postage!!!

Remember the mean, median and mode ?

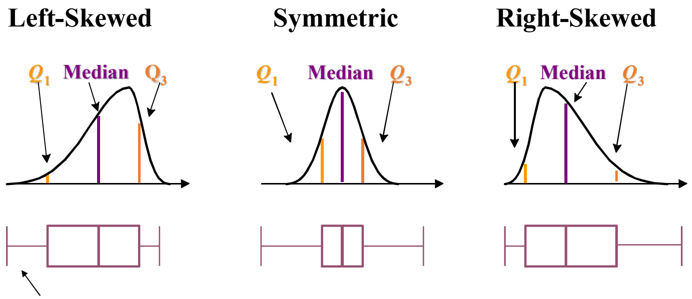
The mean is the average data value,



The value of the mean is strongly affected by skewness and outliers, - more so than the median.

Shape & Box Plot

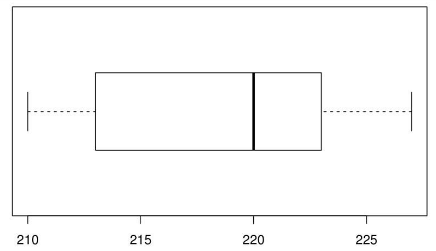
These shapes can also be seen in the boxplots



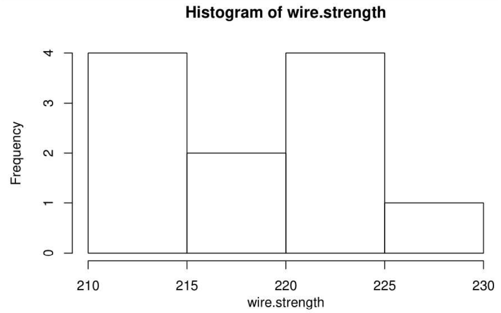
Left skewed - Longer tail on left than right, median may not be central in the box.

Graphical Summary in R: Vector

```r
boxplot(wire.strength, horizontal=TRUE)
```



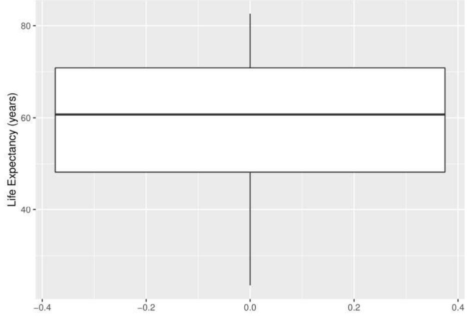
```r
hist(wire.strength)
```



Graphical Summary in R: Dataframe

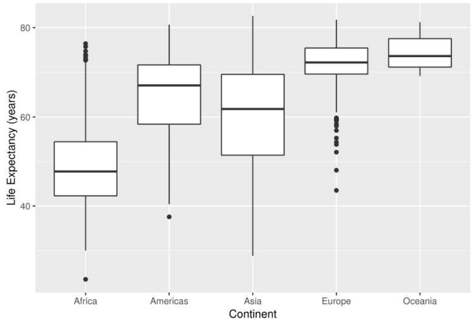
Plot the **boxplot** of life expectancy for gapminder data:

```r
ggplot(gapminder, aes(y = lifeExp)) +
  geom_boxplot() +
  ylab("Life Expectancy (years)")
```

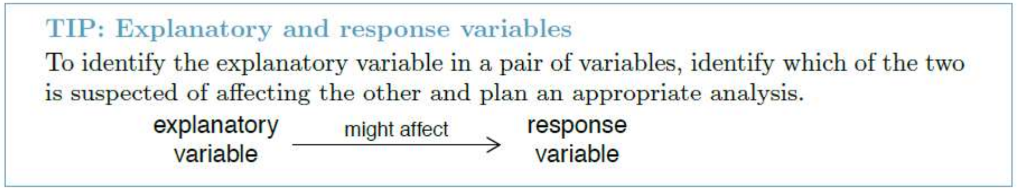


Plot the **boxplot** of life expectancy for different continents:

```r
ggplot(gapminder, aes(x = continent, y = lifeExp)) +
  geom_boxplot() +
  ylab("Life Expectancy (years)") +
  xlab("Continent")
```



Explanatory and response variables

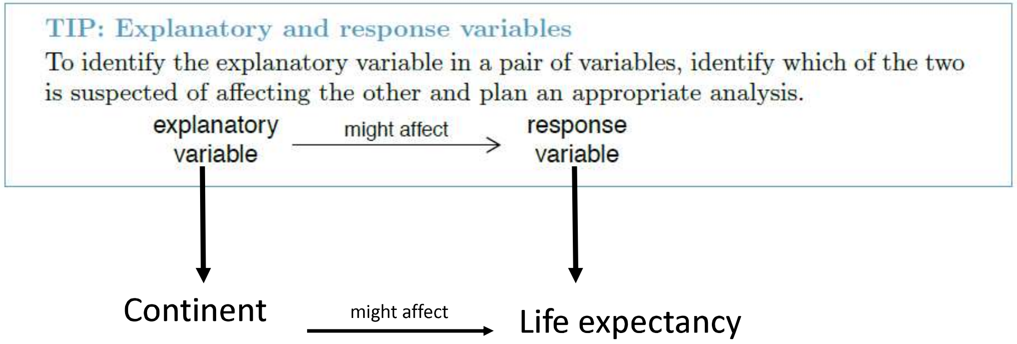


TIP: Explanatory and response variables

To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other and plan an appropriate analysis.

explanatory variable $\xrightarrow{\text{might affect}}$ response variable

Explanatory and response variables



TIP: Explanatory and response variables

To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other and plan an appropriate analysis.

Graphical summaries of data

- Depends on the variable of interest
- Categorical response variable: barchart (n or %) or pie chart
- Categorical response variable with an explanatory variable: grouped barchart
- Continuous response variable: histogram, boxplot, density plot
- Continuous response variable with an explanatory variable: grouped boxplot

Using R

- R statistical computing and visualisation software
- Free open source package,
- Commonly used software for statistics

- 18,000+ contributed packages / libraries
- Lots of tutorials online
- Lots of sources of online help

DataCamp

Search

THE SMARTEST WAY TO

Learn Data Science Online

The skills people and businesses need to succeed are changing. No matter where you are in your career or what field you work in, you will need to understand the language of data. With DataCamp, you learn data science today and apply it tomorrow.

DataCamp

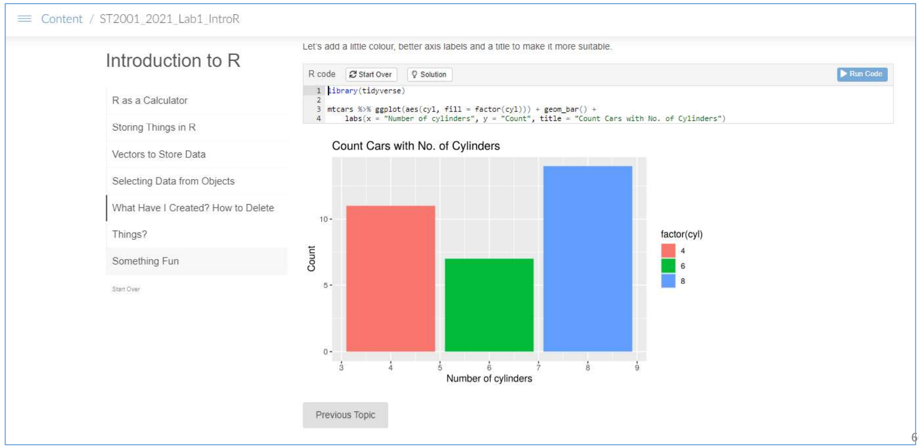
Search

COURSES

- Introduction to Python
- Introduction to R
- Introduction to SQL
- Data Science for Everyone
- Introduction to Data Engineering
- Introduction to Deep Learning in Python

See all courses (338)

A Gentle Start in R



Content / ST2001_2021_Lab1_IntroR

Introduction to R

- R as a Calculator
- Storing Things in R
- Vectors to Store Data
- Selecting Data from Objects
- What Have I Created? How to Delete Things?
- Something Fun

Let's add a little colour, better axis labels and a title to make it more suitable.

```r
library(tidyverse)

mtcars %>% ggplot(aes(cyl, fill = factor(cyl))) + geom_bar() +
    labs(x = "Number of cylinders", y = "Count", title = "Count Cars with No. of Cylinders")
```

cran.r-project.org

+

www.rstudio.com/download

2.1 What are R and RStudio?

moderndive.com

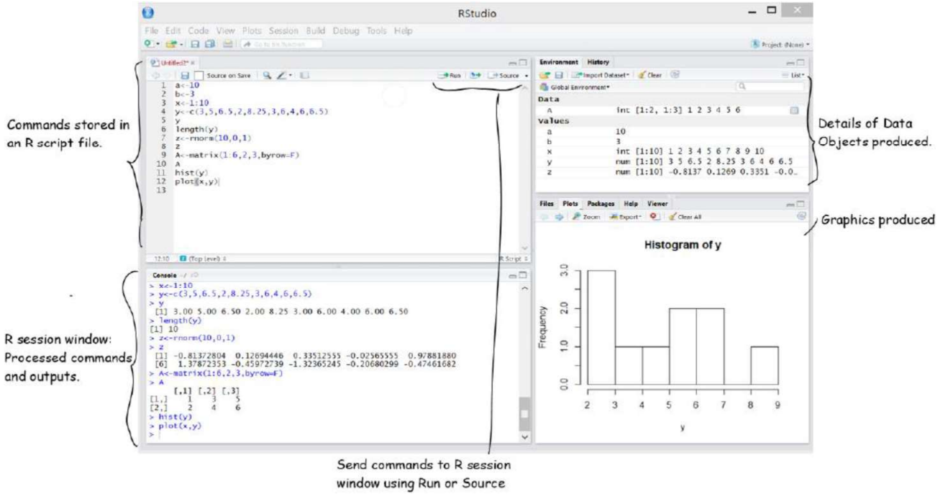
For much of this book, we will assume that you are using R via RStudio. First time users often confuse the two. At its simplest:

- R is like a car's engine
- RStudio is like a car's dashboard

R: Engine	RStudio: Dashboard

More precisely, R is a programming language that runs computations while RStudio is an *integrated development environment (IDE)* that provides an interface by adding many convenient features and tools. So the way of having access to a speedometer, rearview mirrors, and a navigation system makes driving much easier, using RStudio's interface makes using R much easier as well.

RStudio® is an integrated development environment for R.



```
a<-10
b<-3
x<-1:10
y<-c(3,5,6.5,2,8.25,3,6,4,6,6.5)
y
length(y)
z<-rnorm(10,0,1)
z
A<-matrix(1:6,2,3,byrow=F)
A
hist(y)
plot(x,y)
```

```
Data
  A        int [1:2, 1:3] 1 2 3 4 5 6
Values
  a        10
  b        3
  x        int [1:10] 1 2 3 4 5 6 7 8 9 10
  y        num [1:10] 3 5 6.5 2 8.25 3 6 4 6 6.5
  z        num [1:10] -0.8137 0.1269 0.3351 -0.0…
```

```
> x<-1:10
> y<-c(3,5,6.5,2,8.25,3,6,4,6,6.5)
> y
 [1] 3.00 5.00 6.50 2.00 8.25 3.00 6.00 4.00 6.00 6.50
> length(y)
[1] 10
> z<-rnorm(10,0,1)
> z
 [1] -0.81372804  0.12694446  0.33512555 -0.02565555  0.97881880
 [6]  1.37872353 -0.45972739 -1.32365245 -0.20680299 -0.47461682
> A<-matrix(1:6,2,3,byrow=F)
> A
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> hist(y)
> plot(x,y)
>
```

Commands stored in an R script file.

Details of Data Objects produced.

Graphics produced

R session window: Processed commands and outputs.

Send commands to R session window using Run or Source

Installing R and RStudio

Tutorial in installing R and RStudio on your computer (and key packages):

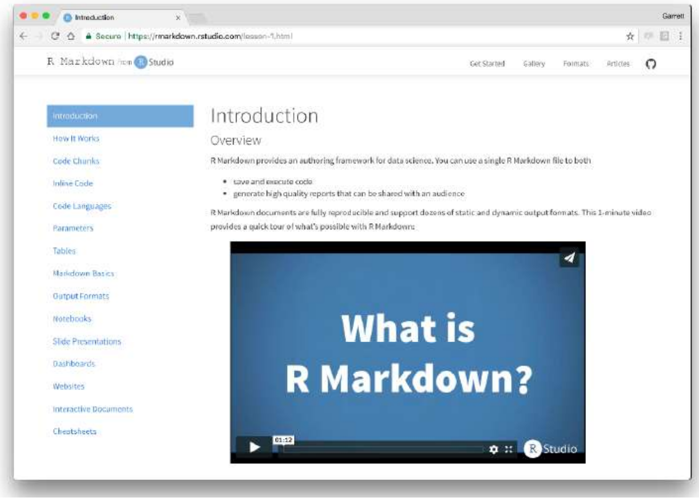
https://jjallaire.shinyapps.io/learnr-tutorial-00-setup/

More instructions videos on Blackboard, but do also google!

Introducing R Markdown

- R Markdown is a file format for making dynamic documents with R
- Written in markdown (an easy-to-write plain text format) and contains:
  - chunks of embedded R code (data management, summaries, graphics, tables, analysis and interpretation)
  - all in the **one** document
- Document can be **knit**ted to html, pdf, word and many other formats!

https://rmarkdown.rstudio.com/lesson-1.html



Introduction

Overview

R Markdown provides an authoring framework for data science. You can use a single R Markdown file to both

- save and execute code
- generate high quality reports that can be shared with an audience

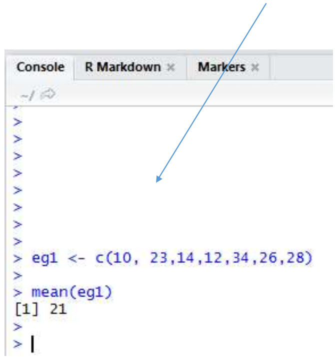
R Markdown documents are fully reproducible and support dozens of static and dynamic output formats. This 1-minute video provides a quick tour of what's possible with R Markdown:

Key Benefits of R Markdown

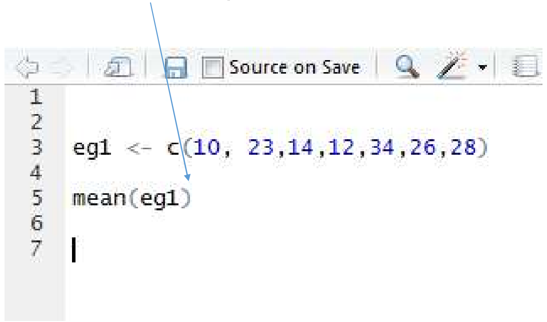
- R Markdown makes it easy to produce statistical reports with code, analysis, outputs and write-up all in one place
- Perfect for reproducible research!
- Easy to convert to different document types

https://github.com/rstudio/cheatsheets/raw/master/rmarkdown.pdf

Drawback of terminal and R script?

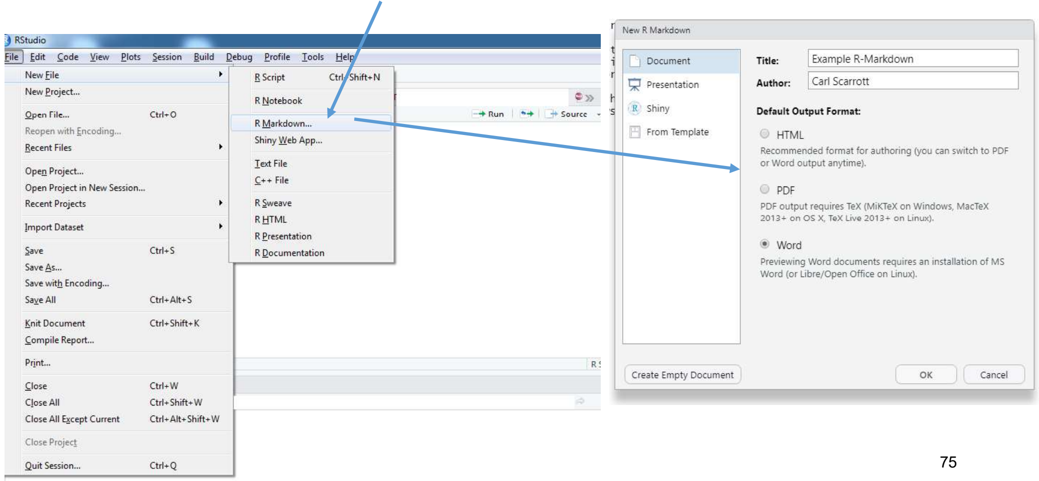


```
>
>
>
>
>
>
>
>
> eg1 <- c(10, 23,14,12,34,26,28)
>
> mean(eg1)
[1] 21
>
> |
```



```


eg1 <- c(10, 23,14,12,34,26,28)

mean(eg1)

|
```

Creating R Markdown Document



Title: Example R-Markdown

Author: Carl Scarrott

Default Output Format:

○ HTML

Recommended format for authoring (you can switch to PDF or Word output anytime).

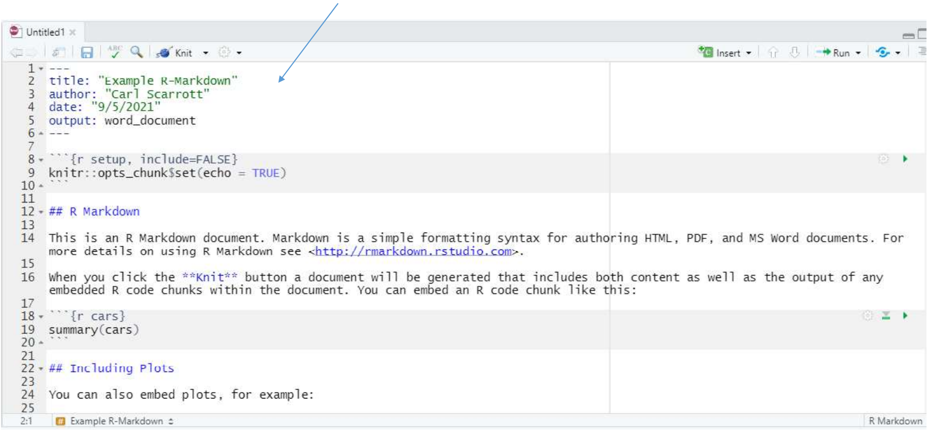
○ PDF

PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

● Word

Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).

Basic R Markdown Document



~~~
---
title: "Example R-Markdown"
author: "Carl Scarrott"
date: "9/5/2021"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For
more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any
embedded R code chunks within the document. You can embed an R code chunk like this:

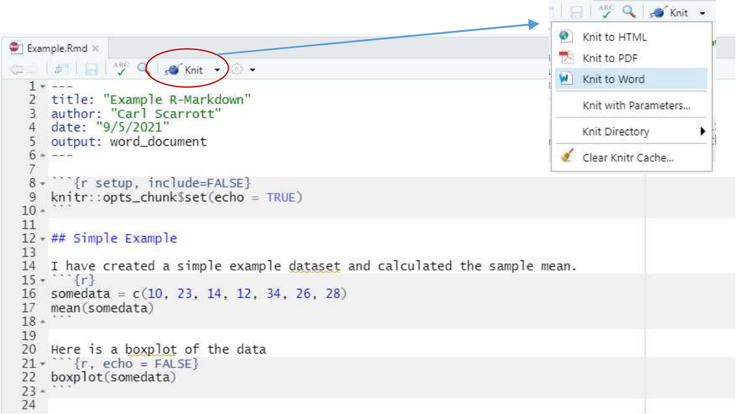
```{r cars}
summary(cars)
```

Including Plots

You can also embed plots, for example:

```
```

Edit and "knit" Document

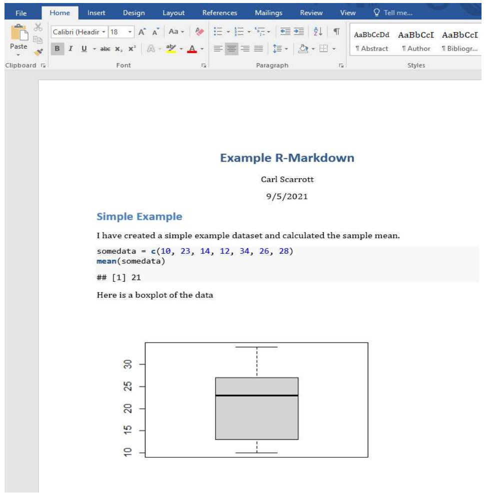


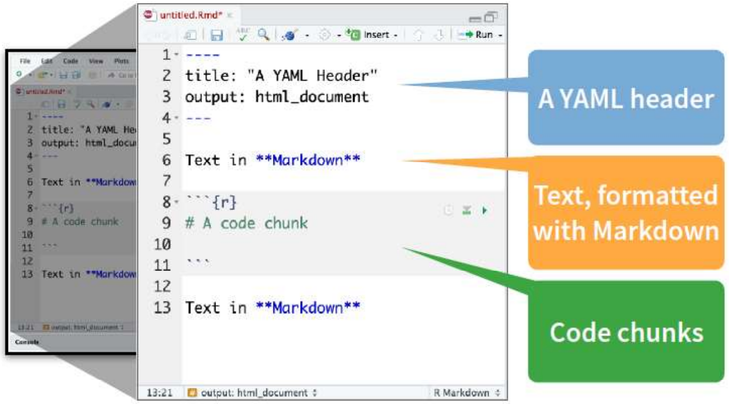
~~~
---
title: "Example R-Markdown"
author: "Carl Scarrott"
date: "9/5/2021"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

Simple Example

I have created a simple example dataset and calculated the sample mean.
```{r}
somedata = c(10, 23, 14, 12, 34, 26, 28)
mean(somedata)
```

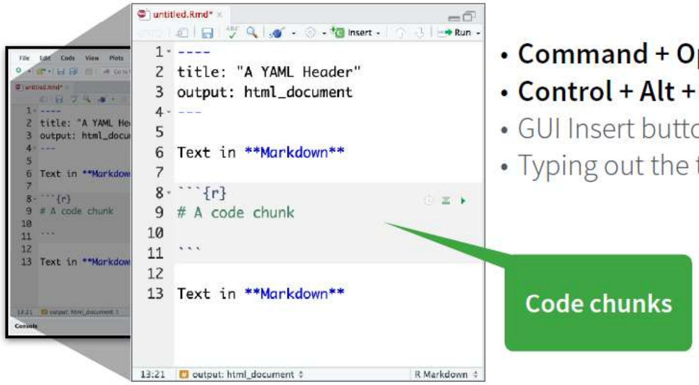
Here is a boxplot of the data
```{r, echo = FALSE}
boxplot(somedata)
```

```


Example.Rmd
Knit
Knit to HTML
Knit to PDF
Knit to Word
Knit with Parameters...
Knit Directory
Clear Knitr Cache...




# R Markdown knitted to Word




File
Home
Insert
Design
Layout
References
Mailings
Review
View
Tell me...
Paste
Calibri (Headir
18
Clipboard
Font
Paragraph
Styles
AaBbCcDd
AaBbCcI
AaBbCcI
¶ Abstract
¶ Author
¶ Bibliogr...


Example R-Markdown

Carl Scarrott

9/5/2021

Simple Example

I have created a simple example dataset and calculated the sample mean.

```
somedata = c(10, 23, 14, 12, 34, 26, 28)
mean(somedata)

## [1] 21
```

Here is a boxplot of the data


10
15
20
25
30




# Structure

R Markdown documents contain **three types of content**



```
---
title: "A YAML Header"
output: html_document

Text in **Markdown**

```{r}
# A code chunk

```

Text in **Markdown**
```


untitled.Rmd*
Insert
Run
13:21
output: html_document
R Markdown
File
Edit
Code
View
Plots
Console
A YAML header
Text, formatted with Markdown
Code chunks




# Code Chunks

Write and execute code in a **chunk**. Insert with



```
---
title: "A YAML Header"
output: html_document

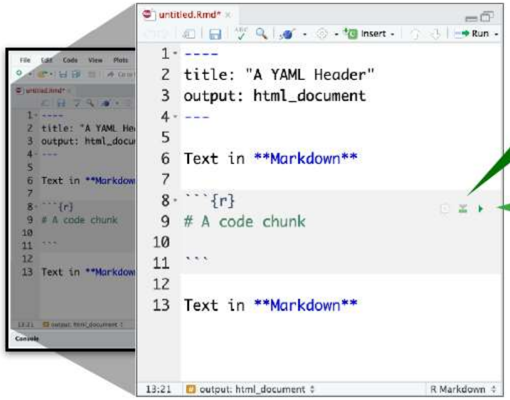
Text in **Markdown**

```{r}
# A code chunk

```

Text in **Markdown**
```

- **Command + Opt + I** ()
- **Control + Alt + I** ()
- GUI Insert button
- Typing out the tick marks
```

Code Chunks

Write and execute code in a **chunk**.



~~~~
----
title: "A YAML Header"
output: html_document
---

Text in **Markdown**

```{r}
# A code chunk

```

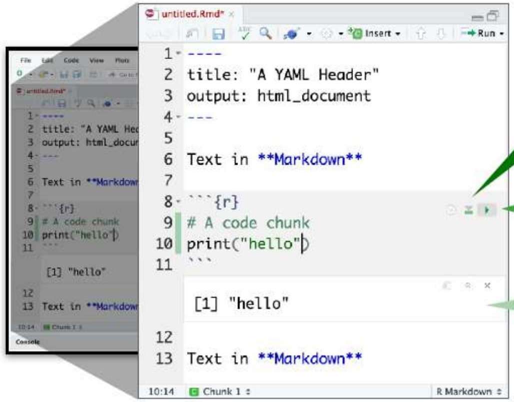
Text in **Markdown**
~~~~

Click to run all code chunks above

Click to run code in chunk

Code Chunks

Write and execute code in a **chunk**.



~~~~
----
title: "A YAML Header"
output: html_document
---

Text in **Markdown**

```{r}
# A code chunk
print("hello")
```

[1] "hello"

Text in **Markdown**
~~~~

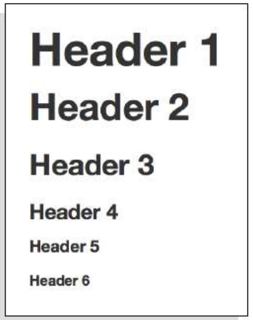
Click to run all code chunks above

Click to run code in chunk

Code result

Headers

```
# Header 1
## Header 2
### Header 3
#### Header 4
##### Header 5
###### Header 6
```



Header 1

Header 2

Header 3

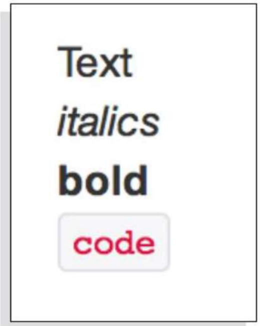
Header 4

Header 5

Header 6

Text

```
Text
_italics_
__bold__
`code`
```



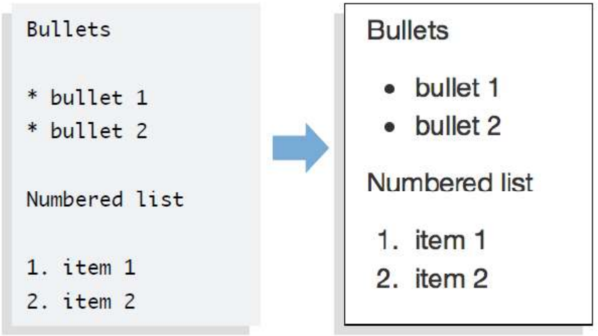
Text

italics

bold

`code`

Lists



```
Bullets

* bullet 1
* bullet 2

Numbered list

1. item 1
2. item 2
```

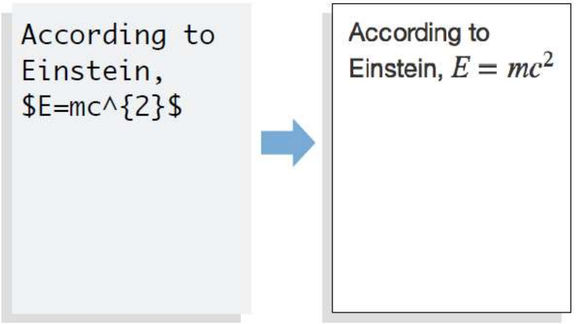
Bullets

- bullet 1
- bullet 2

Numbered list

1. item 1
2. item 2

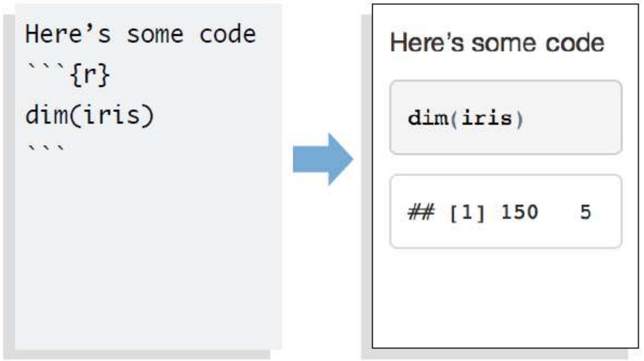
Equations



```
According to
Einstein,
$E=mc^{2}$
```

According to Einstein, $E = mc^2$

Code chunks



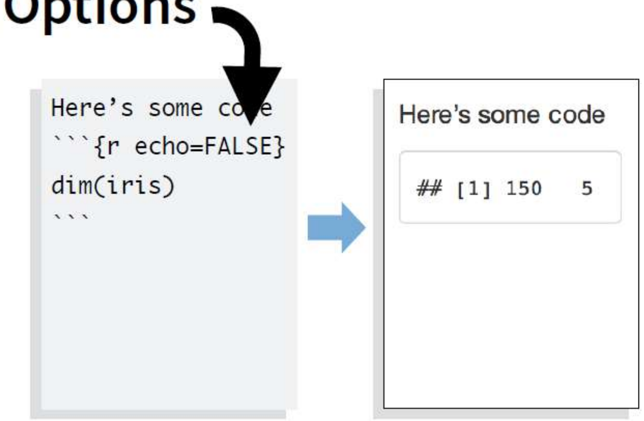
~~~
Here's some code
```{r}
dim(iris)
```
~~~

Here's some code

```
dim(iris)
```

```
## [1] 150   5
```

Chunk Options

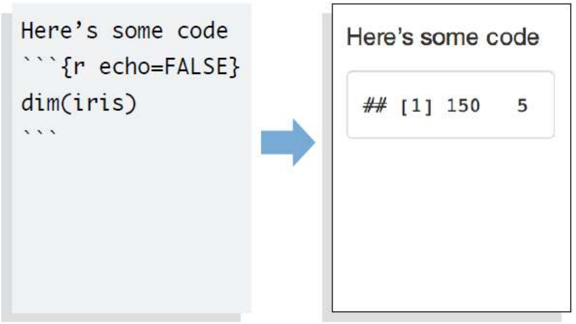


~~~
Here's some code
```{r echo=FALSE}
dim(iris)
```
~~~

Here's some code

```
## [1] 150   5
```

echo = FALSE



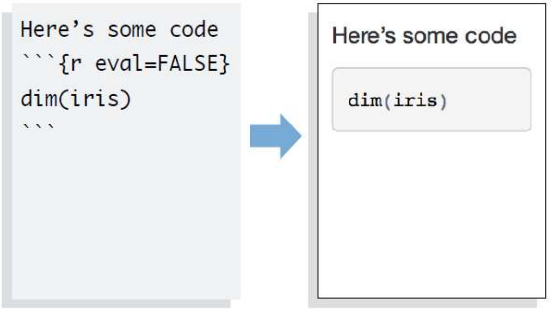
~~~~
Here’s some code
```{r echo=FALSE}
dim(iris)
```
~~~~

Here’s some code

```
## [1] 150    5
```

Displays code results, but **not code**

eval = FALSE



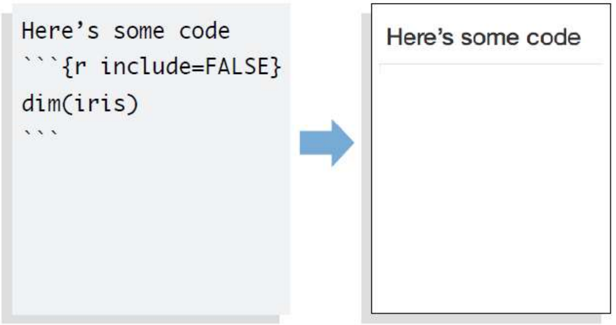
~~~~
Here’s some code
```{r eval=FALSE}
dim(iris)
```
~~~~

Here’s some code

```
dim(iris)
```

Displays code, but **not results** (code is not run)

include = FALSE

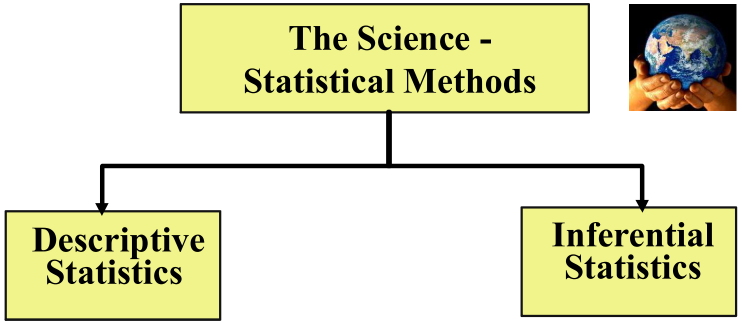


~~~~
Here’s some code
```{r include=FALSE}
dim(iris)
```
~~~~

Here’s some code

Displays **neither code not results** (but code is run)

Statistical Methods



The Science - Statistical Methods

Descriptive Statistics

Inferential Statistics

Inferential Statistics: *science of using the **information in your sample** to say (i.e. to “**infer**”) something **about the population** of interest*

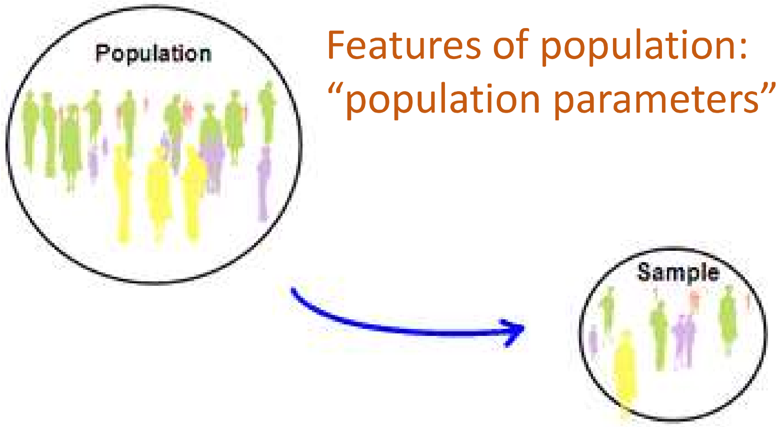
More precisely, to infer or **estimate the value of population parameters** using **sample statistics as estimates**

We will come back to this later!

Descriptive Statistics: *Science of summarizing data, numerically and graphically…*

Analysis methods applicable depends on the variable being measured and the research questions which you are trying to answer …

How would you collect the data?

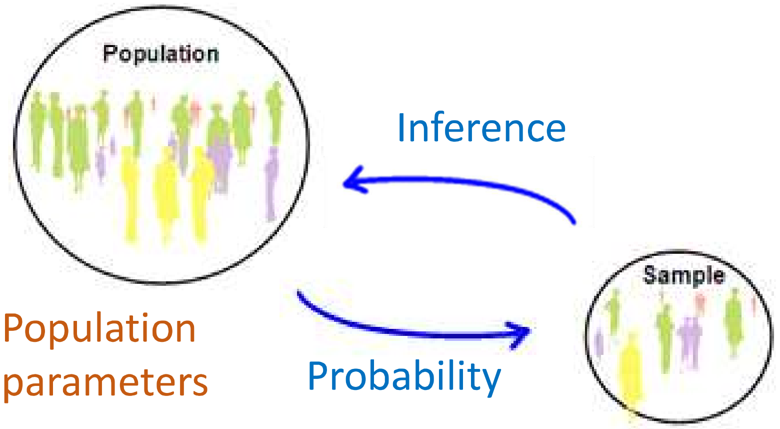


Features of population:
“population parameters”

Estimate/Infer from sample:
“sample statistics”

Key concepts

- ***Population***
  - ***A parameter*** is a single value summarising some feature of variable of interest in the *population*
  - It is usually unknown…
- ***Sample***
  - ***A statistic*** is a single value summarising the observed values of the variable from the *sample* collected
  - Sample statistics will vary from sample to sample
  - Source of uncertainty….



Inference

Population parameters

Probability

Sample statistics

Inference / Inferential Statistics :
estimating a parameter using sample statistic

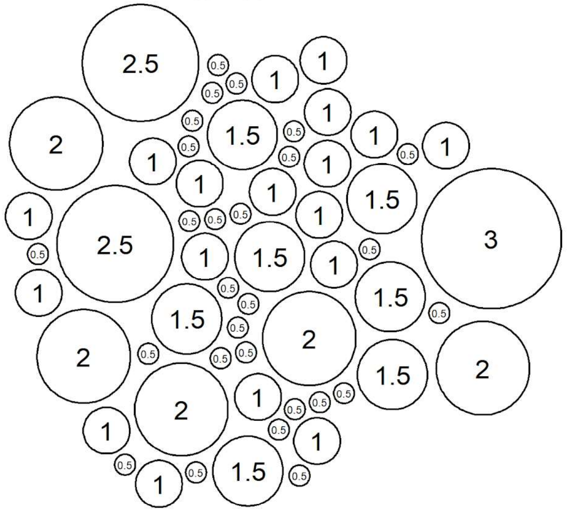
Inferential Statistics:

Inference is the process of making decisions about a population based on information in a sample

Science of Inferential Statistics:

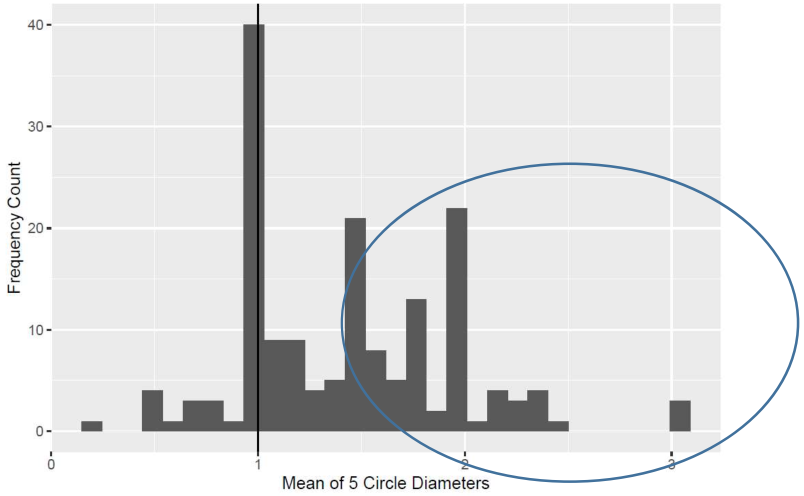
to **infer or estimate population parameters** using **sample statistics**

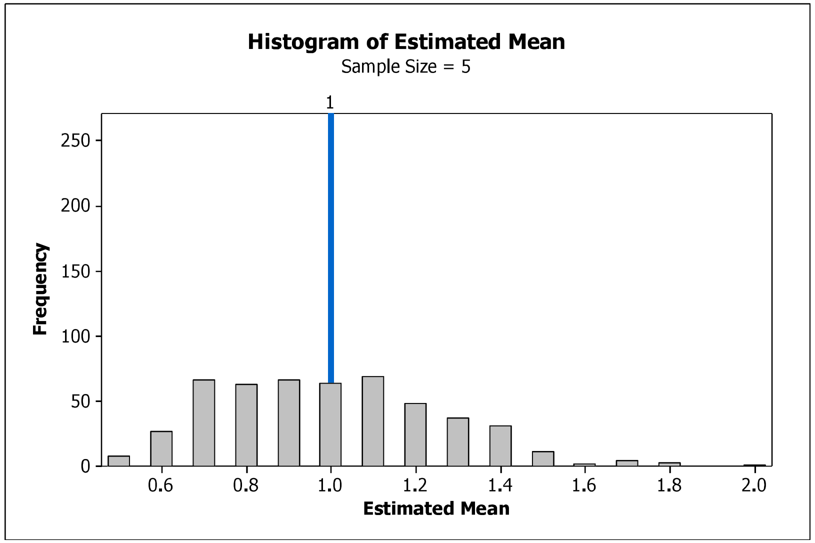
Population of Circles (Diameters)
N = 60

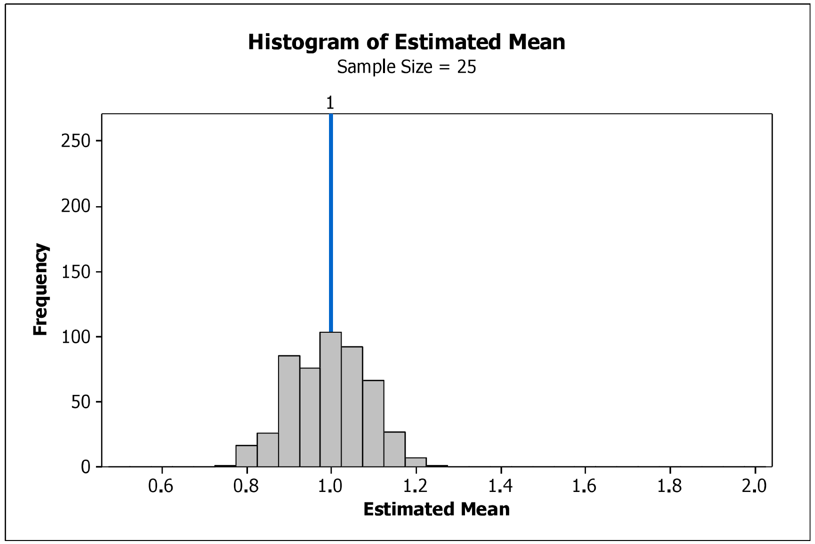


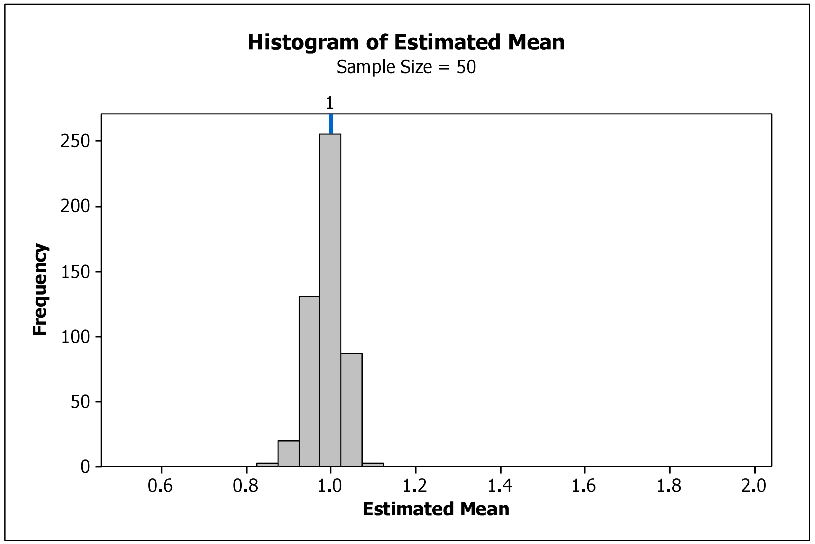
Take a representative sample of *5* circles from the population of *60* circles and use the sample mean as an estimate of the true population mean

Your Samples





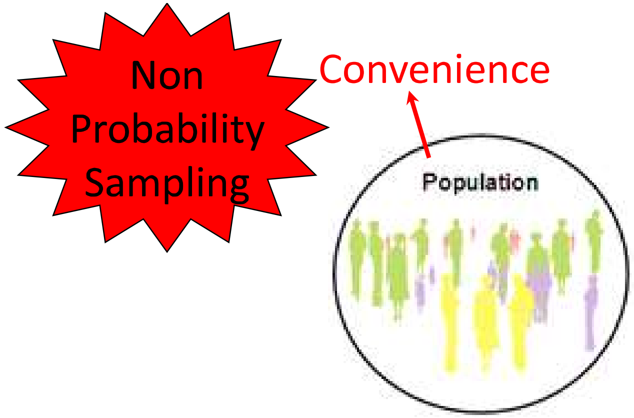




A consequence of **natural variation** is that two samples drawn from the same population will usually give different estimates of the population parameters

Referred to as **sampling variation**

How can we choose a representative sample of size 500 University of Galway students?


Non Probability Sampling

Convenience

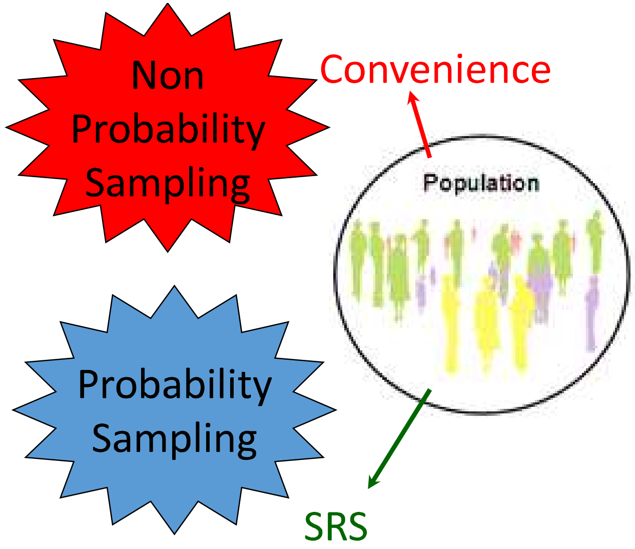
Non-probabilistic sampling methods are techniques of obtaining a sample that is not chosen at random and may be subject to **sampling bias**.

Examples of Convenience Sampling

- In the ***amount spent on rent weekly*** problem sending emails to students in the hope of a response may give this result
- Could sample by standing at entrance to college bar or on concourse **–** another example of convenience sampling

An issue with Convenience Sampling

- Leave it to experimental unit to choose to complete a survey or opinion poll
  - A response may be more likely to be received because a responder has a particularly strong opinion
  - If sample consists of mainly such strong opinions, then sample may not be representative of population


Non Probability Sampling

Convenience

Probability Sampling

SRS

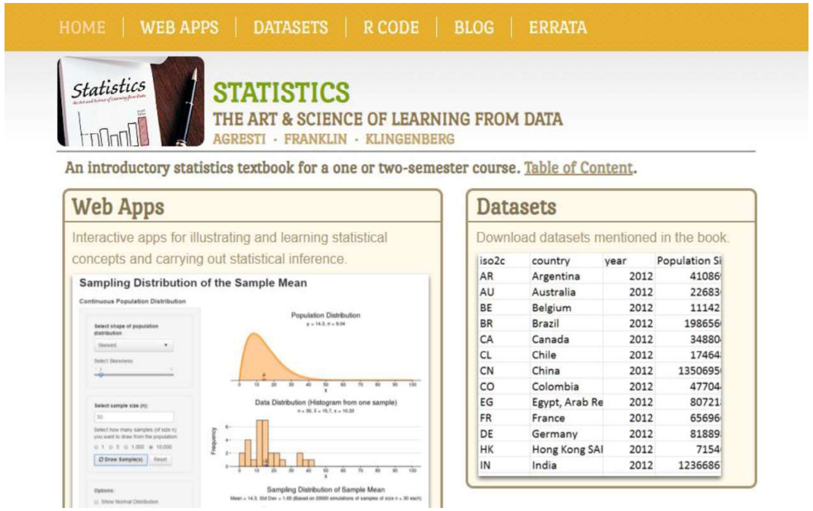
Simple Random Sampling

”Subset of n units chosen from population of N units, chosen randomly so that every unit has same chance of selection”

Or equivalently, every randomly subset of same size has same change of being observed

Sound easy…

Just “table and label”!


HOME | WEB APPS | DATASETS | R CODE | BLOG | ERRATA
Statistics
Sampling Distribution of the Sample Mean
Continuous Population Distribution
Population Distribution
Data Distribution (Histogram from one sample)
Frequency
Sampling Distribution of Sample Mean

STATISTICS
THE ART & SCIENCE OF LEARNING FROM DATA
AGRESTI · FRANKLIN · KLINGENBERG

An introductory statistics textbook for a one or two-semester course. Table of Content.

Web Apps

Interactive apps for illustrating and learning statistical concepts and carrying out statistical inference.

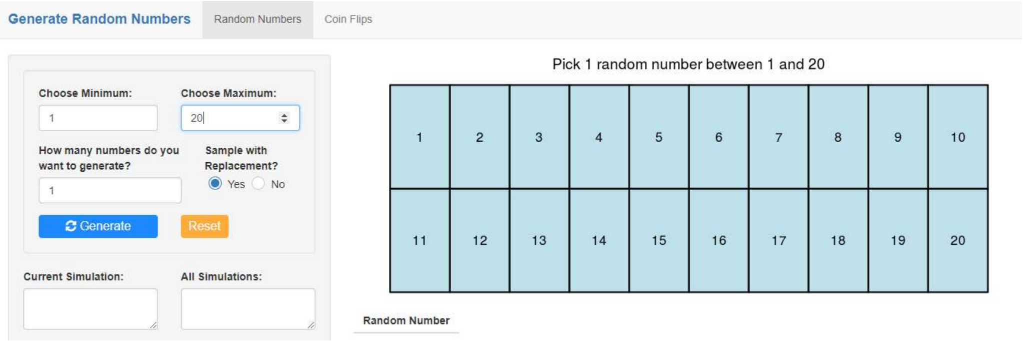
Datasets

Download datasets mentioned in the book.

iso2c	country	year	Population Si
AR	Argentina	2012	41086
AU	Australia	2012	22683
BE	Belgium	2012	11142
BR	Brazil	2012	198656
CA	Canada	2012	34880
CL	Chile	2012	17464
CN	China	2012	1350695
CO	Colombia	2012	47704
EG	Egypt, Arab Re	2012	80721
FR	France	2012	65696
DE	Germany	2012	81889
HK	Hong Kong SAI	2012	7154
IN	India	2012	1236686

https://www.artofstat.com

https://istats.shinyapps.io/RandomNumbers/



Choose Minimum: 1

Choose Maximum: 20

How many numbers do you want to generate? 1

Sample with Replacement? ● Yes ○ No

Current Simulation:

All Simulations:

Pick 1 random number between 1 and 20

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20

Simple **random** sample

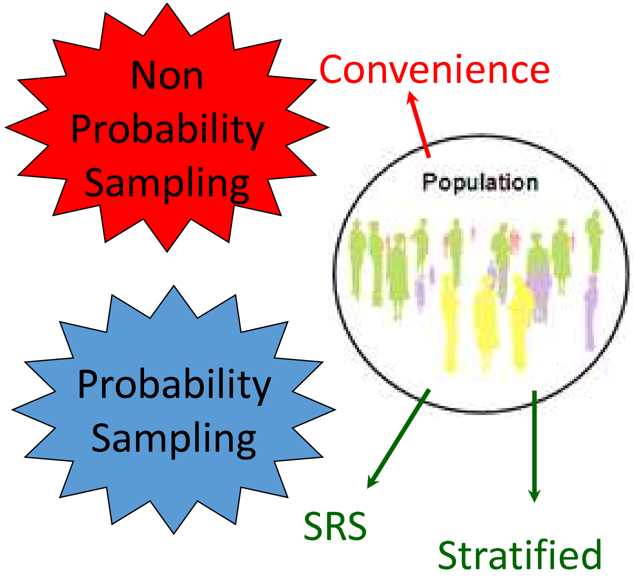
Difficulties:

- Obtaining a sampling frame (list of all experimental units)
- Possibly time consuming / expensive
- Minority groups, by chance, may not be represented in sample

e.g. population of N = 17,520 University of Galway students,

400 of which are mature students

how many in sample of n = 500?



Non Probability Sampling

Convenience

Probability Sampling

SRS

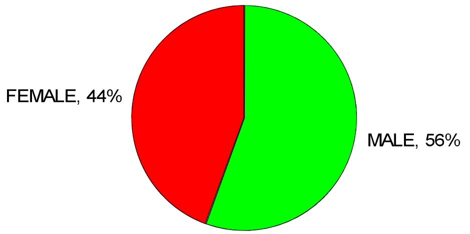
Stratified

Stratified Random Sampling

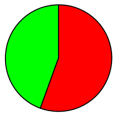
(i) Split entire population into homogeneous groups, called strata

(ii) Take a SRS from each stratum

Proportional Allocation



N = 10,000



n = 100

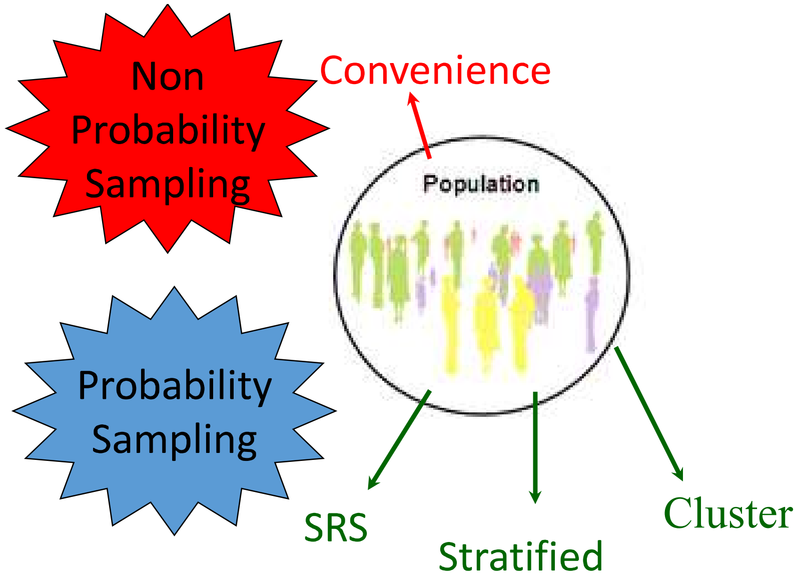
Stratified Compared to SRS

- Ensure representation from minority groups
- Estimates of the population parameters per strata may be of interest
- Possibly reduction in cost per observation in the survey
- Increased accuracy as reduced sampling error (less variation within a stratum)

Stratified Compared to SRS

- Can you correctly allocate each individual to one and only one stratum?
- Should every group receive equal weight?
- What if some strata are more varied than others?
- Take into account mean, variance and cost to get "optimal allocation"

What if a sampling frame (or strata criteria) is unavailable?



Non Probability Sampling

Convenience

Probability Sampling

SRS

Stratified

Cluster

Cluster Sampling

Instead of randomly choosing individuals,

a SRS of collections or groups of individuals is taken

Cluster Sampling

Population is broken up into regions or groups, usually a natural partition, called a ***cluster***

e.g., geographical areas, or a class!

- Clusters are assumed representative of entire population

  (internally heterogenous, between are homogeneous)

- Small number of clusters are selected at random
- ***Every* individual** within a cluster are observed

Cluster over Stratified

- Sampling frame not necessarily needed
- May be more practical and / or economical than SRS or stratified sampling
- Will be biased if entire cluster not sampled
- Careful if homogeneity within cluster and heterogeneity between clusters as this can increase sample error

Note:

in stratified sampling all strata are sampled while in cluster sampling only some clusters are sampled

This is a crucial point!

Summary

- Try to estimate population parameters with sample statistics
- Want sample to be as representative of the population as possible

- Probability based sampling schemes are best in terms of minimising chance of bias

Typical Exam Questions

1. Which of the following statements is true regarding a population:

   a) it must refer to people;
   b) it is a collection of individuals or objects;
   c) neither of the above.

Typical Exam Questions

2. A sampling frame is:

   a) the list of units from which the sample is chosen;
   b) a table of random numbers;
   c) a non-probabilistic sampling method.

Typical Exam Questions

3. Sampling that divides the population in subgroups and chooses a proportionate number from each subgroup at random is called:

   a) cluster sampling;
   b) quota sampling;
   c) stratified sampling.

Collecting the Data (Sampling)

- Observational Study

- Designed Experiment

Observational studies & experiments

- **Observational** study:
  - data collected only by observing what occurs (e.g. surveys, historical records)
- When researchers want to investigate **causal relationships** best to conduct an **experiment**
  - Usually there will be both an explanatory and a response variable
- Be wary of confounding variables.

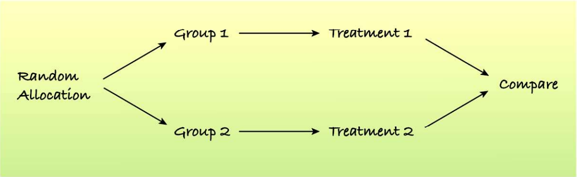
Designed (comparative) Study

- An experiment allows us to prove a cause-and-effect relationship
- Experimenter must identify:
  - at least one explanatory variable, called a factor, to manipulate; and
  - at least one response variable to measure
- They must control any other nuisance factors that could influence the response, e.g. weather, day of week, …

Designed (comparative) Study

- An experiment allows us to prove a cause-and-effect relationship
- An experiment will:
  - *Define* treatment factor(s)
  - *Randomly* assigns subjects to treatment levels
  - *Compares* responses of the subject groups across treatment levels
- Key difference between an observational study and an experiment is that in an experiment we apply a treatment to the subjects in a controlled way

Comparative Studies (Independent Samples)



Random Allocation

Group 1

Treatment 1

Group 2

Treatment 2

Compare

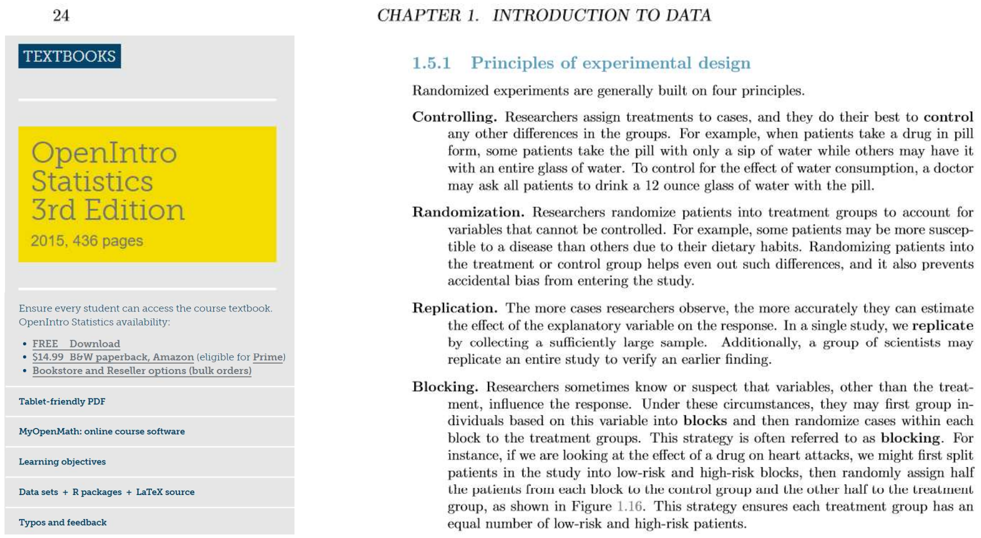
- Randomisation
- Control Group
- Baseline (Pre) measurement
- Blinding
- Replication, not pseudo-replicates

Designed Study (cont.)

- When humans are involved, the term “experimental units” is commonly replace with subjects or participants
- A treatment is a combination of specific levels from all the factors that an experimental unit receives
- A baseline measurement is the initial measurements of the response at the beginning of the experiment
- Changes from the baseline due to varying the treatment level are of usually interest
- A control group receive a standard treatment (usually a placebo, or no treatment at all) called a control treatment, often the response is not expected to change from baseline for the control group

Principles of Study Design

- Controlling (not just control group)
- Randomisation
- Replication
- Blocking (stratifying)
- Blinding



24

TEXTBOOKS

OpenIntro Statistics 3rd Edition

2015, 436 pages

Ensure every student can access the course textbook. OpenIntro Statistics availability:

- FREE Download
- $14.99 B&W paperback, Amazon (eligible for Prime)
- Bookstore and Reseller options (bulk orders)

Tablet-friendly PDF

MyOpenMath: online course software

Learning objectives

Data sets + R packages + LaTeX source

Typos and feedback

CHAPTER 1. INTRODUCTION TO DATA

1.5.1 Principles of experimental design

Randomized experiments are generally built on four principles.

Controlling. Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups. For example, when patients take a drug in pill form, some patients take the pill with only a sip of water while others may have it with an entire glass of water. To control for the effect of water consumption, a doctor may ask all patients to drink a 12 ounce glass of water with the pill.

Randomization. Researchers randomize patients into treatment groups to account for variables that cannot be controlled. For example, some patients may be more susceptible to a disease than others due to their dietary habits. Randomizing patients into the treatment or control group helps even out such differences, and it also prevents accidental bias from entering the study.

Replication. The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we **replicate** by collecting a sufficiently large sample. Additionally, a group of scientists may replicate an entire study to verify an earlier finding.

Blocking. Researchers sometimes know or suspect that variables, other than the treatment, influence the response. Under these circumstances, they may first group individuals based on this variable into **blocks** and then randomize cases within each block to the treatment groups. This strategy is often referred to as **blocking**. For instance, if we are looking at the effect of a drug on heart attacks, we might first split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and the other half to the treatment group, as shown in Figure 1.16. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

Principles of Experimental Design

- Controlling:
  - Need to control all nuisance factors that may influence response or sources of variation, other than the treatments of interest by making conditions as similar as possible for all treatment groups
- Randomisation:
  - Attempts to “equalise” the effects of unknown or uncontrollable sources of variation
    - It does not eliminate the effects of these sources, but it tries to minimise their impact
  - Without randomisation, you do not have a valid experiment and will not be able to use the powerful methods of statistics to draw conclusions from your study

Principles of Experimental Design

- Replication:
  - Repeat the experiment, applying the treatments to a number of subjects
- Blocking (stratifying):
  - Sometimes attributes of the experimental units, that we are not studying, and that we can't control may affect the outcomes of an experiment
  - Solution: group similar individuals together and then randomise within each of these blocks, to remove much of variability between the blocks
  - Note: blocking is an important compromise between randomisation and control but is not *required* in an experimental design

Blocking

- When experimental units within a group are similar, it's often a good idea to gather them together into blocks
  - Blocking isolates the variability due to the differences between the blocks so that we can see the differences due to the treatments more clearly
  - Basically it removes a source of noise so signal stronger
- When randomization occurs only within the blocks, we call the design a randomized block design
- Blocking is same idea for experiments as stratifying is for sampling

Principles of Experimental Design

- Blinding:
  - Refers to the concealment of treatment allocation, from one or more individuals involved in a study
    - Although randomisation minimizes differences between treatment groups at the start of the study, it does nothing to prevent differential treatment of the groups during trial, or the differential assessment of outcomes, which may result in biased estimates of treatment effects
  - Best practice to minimise the likelihood of differential treatment or assessments of outcomes is to blind as many individuals as possible in a trial (e.g. participants, experimenters, statisticians)

Blinding

- There are two main classes of individuals who can affect the outcome:
  - those who could influence the treatment response (usually the subjects, treatment administrators or technicians)
  - those who evaluate the results (statisticians, researchers, physicians, etc.)
- When all individuals in *either one* of these classes are blinded, an experiment is said to be single-blind (usually the first class)
- When everyone in *both* classes is blinded, the experiment is called double-blind

Placebos

- Often simply applying *any* treatment can induce an improvement
- To separate out the effects of the treatment of interest, we can use a control treatment that mimics the treatment itself
- A “fake” treatment like the treatment being tested is called a placebo (e.g. saline solution or inert pill)

- Placebos are the best way to blind subjects from knowing whether they are receiving the treatment or not

Placebo Effect

- A placebo effect occurs when taking the sham treatment results in a change in the response variable
- Just being involved in an experiment can change behavior or feelings

- Highlights both the importance of effective blinding and the importance of comparing treatments with a control
- Placebo controls are so effective they should be considered an essential tool for blinding whenever possible

Best Practice for Experiments

- Usually:
  - randomised
  - comparative
  - double-blind
  - have control group (either placebo or a standard treatment)

Observational Studies

In an observational study we may compare units that **happened** to receive different treatments

Example: **Smoking & Lung Cancer**

- Simply comparing across groups that smoked or not
- No control of treatment allocation
- Are those prone to smoking also naturally prone to lung cancer?
- Could identifying “possible” causes, but cannot establish causation

Only properly designed and executed experiments can reliably demonstrate cause-and-effect

What Can Go Wrong?

- Don't give up if you can't run an experiment
  - If we can't perform an experiment, an observational study may be good option
- Beware of confounding; some other unmeasured variables that has an effect on the response variable intertwined in the experiment (e.g. severity of disease at baseline)
- Use randomisation whenever possible to minimise risk of confounding
- Always report any possible unavoidable confounding

What Can Go Wrong?

- Bad things can happen even to good experiments
  - Protect yourself by recording additional information
  - Account for nuisance factors in your modelling, even if you randomised
- Don't spend your entire budget on the first run
  - Run a small pilot experiment first
  - You may learn some things that will help you make the full-scale experiment better

Proposed Design

Pilot Study

Summary

- Observational studies are tricky to analyse
- Experimental studies are the key to establishing cause and effect
- Both observational and experimental studies need randomisation to collect unbiased data. But they do so in different ways and for different purposes:
  - Observational studies attempt to randomly select participants from the population.
  - Experiments are usually done by randomly assigning the treatments to the experimental units (e.g. patients) to reduce bias.
- No Control no Experiment
- Carry out a Pilot study
- Consult your favorite Statistician (bring gifts)

Topic 4: Probability

First, an introduction: Dr Nicola Fitz-Simon

- Studied statistics at TCD, PhD (2006)
- Worked as a statistician on research studies and lecturing in the UK at Oxford University, LSHTM, Imperial College London
- Most recent post in Galway in the Clinical Research Facility
- Research interests in statistical methods for causal inference
- Contact details next week ...

Summary so far

- Statistical inference involves sampling from populations to generate an estimate (i.e. a statistic) of a population parameter of interest.
- Choosing a sample at *random* is crucial. Subjective sampling will lead to bias (e.g. circles example).

The challenge …

- It is well and good to see what happens if you take lots of samples at random from a population.
- In practice you will only be taking one sample !
- What can you say about how likely your statistic is to be a good guess of the population parameter of interest ?
- To answer this we need to look at some probability theory.

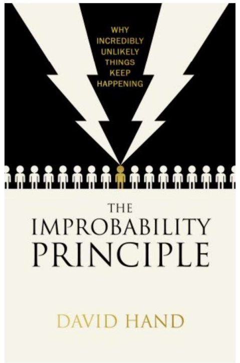
The Role of Probability

- Probability provides the **framework** for the study and application of statistics.

- Probability concepts will be introduced in the next few lectures.

Learning Objectives

1. Interpret probabilities and calculate probabilities of events
2. Calculate the probabilities of joint events
3. Interpret and calculate conditional probabilities
4. Determine independence and use independence to calculate probabilities
5. Understand Bayes' theorem and when to use it



'A lifetime's worth of wisdom'
Steven D. Levitt, co-author of Freakonomics
The International Bestseller
Thinking, Fast and Slow
Daniel Kahneman
Winner of the Nobel Prize

Probability:

How likely .... ?

Tossing a coin

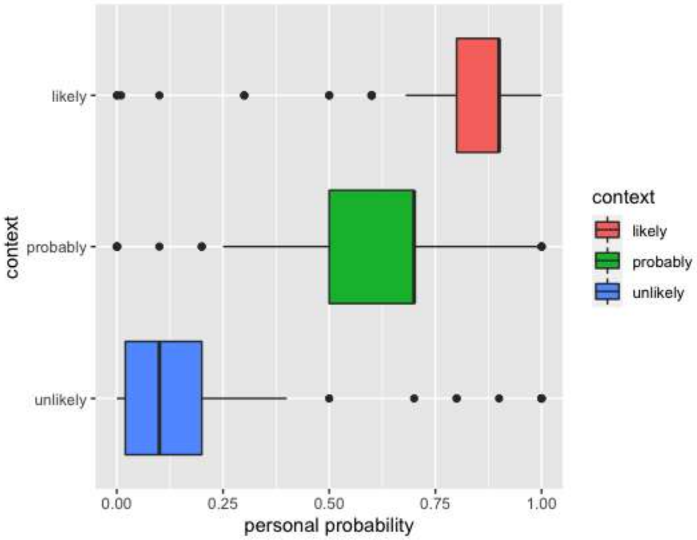
- If I toss a coin, what is the probability it will turn up heads?

Tossing a coin

Magician and statistician Persi Diaconis found that when tossing a coin and catching it in the hand the probability of the same face turning up as initially is about 0.51. https://www.youtube.com/watch?v=AYnJv68T3MM

What are probabilities?

- 6-sided die about to be tossed for the first time
- **Classical**: 6 possible outcomes, by symmetry each equally likely to occur
- **Frequentist**: Empirical evidence shows that similar dice thrown in the past have landed on each side about equally often
- **Subjective**: the degree of individual belief in occurrence of an event – can be influenced by classical or frequentist arguments, eg here may be willing to bet at a rate of 1/6 on any side
- Subjective probabilities also influenced by other reasons when symmetry arguments don't apply and repeated trials are not possible



Probability

- The probability of an event **A** is the number of (equally likely and disjoint) outcomes in the event divided by the total number of (equally likely and disjoint) possible outcomes.

$$P(\mathbf{A}) = \frac{\text{\# of outcomes in } \mathbf{A}}{\text{\# of possible outcomes}}$$

$(0 \leq \mathrm{P(A)} \leq 1)$ **

All possible outcomes (Sample spaces)

- The set of all possible outcomes of a random experiment is called the sample space, *S*.
- *S* is discrete if it consists of a finite or countable infinite set of outcomes.
- *S* is continuous if it contains an interval of real numbers.
- P(S)=1 **

Examples of Sample Spaces

- Toss a coin twice
  S = { HH, HT, TH, TT }
- Roll a pair of dice and record numbers
  S = {(1,1),(1,2),…,(1,6),(2,1),…, (2,6),…,(6,6)}
- Roll a pair of dice and record total score
  S = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}
- Toss a coin until first tail appears
  S = {T, HT, HHT, HHHT, …}
- Measure duration of charge of mobile phone battery
  $S = \{ t \mid t \geq 0 \}$

Events

An **Event** is a specific collection of sample points / possible outcomes.

An event is denoted by **E** or capital letters at the start of the alphabet, A, B, C etc.

Events

An **Event** is a specific collection of sample points / possible outcomes. An event is denoted by **E**

A **Simple Event** is a collection of only one sample point/possible outcome

- Eg: Throw a die – Event 1 : get a 4

  E_1={4} - a simple event

Events

An **Event** is a specific collection of sample points / possible outcomes. An event is denoted by **E**

A **Simple Event** is a collection of only one sample point/possible outcome

- Eg: Throw a die – Event 1 : get a 4

  E_1={4} - a simple event

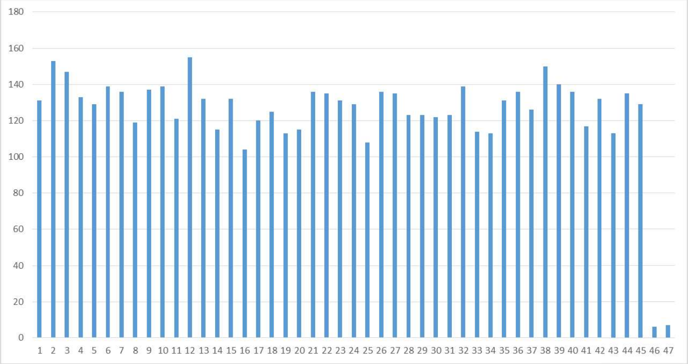
A **Compound Event** is a collection of more than one sample point/possible outcomes

- Eg: Throw a die – Event 2: get at least a 4

  E_2={4,5,6} - a compound event

Example: Lottery

National Lottery



Since launch of Lotto 6/45 on November 4th 2006.
Balls 46 and 47 introduced September 3rd 2015.

Learning to count ..

- How many outcomes are there for the lotto when the outcome of interest is to guess 6 numbers correctly from 47 ?
- How many ways can this occur ?
- Combinatorics: permutations and combinations

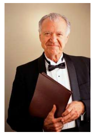
To calculate a probability:

If each sample point in the sample space is equally likely

$$P(\text{Event}) = \frac{\textit{Number of outcomes in event}}{\textit{Total number outcomes in sample space.}}$$

So we need to learn how to count….

Counting by Multiplication

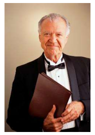


The fixed-price dinner at a restaurant provides the following choices:

Appetizer: Soup or Salad
Main Course: Baked chicken,
Broiled beef patty,
Baby beef liver,
or Roast beef
Dessert: Ice-cream or Cheese cake

How many different three course meals can be ordered?

Counting by Multiplication



The fixed-price dinner at a restaurant provides the following choices:

Appetizer: Soup or Salad
Main Course: Baked chicken,
Broiled beef patty,
Baby beef liver,
or Roast beef
Dessert: Ice-cream or Cheese cake

How many different three course meals can be ordered? 2x4x2=16

The multiplication principle

If a task consists of a sequence of choices in which there are
p selections for the first choice,
q selections for the second choice,
r selections for the third choice,
and so on,
then the task of making these selections can be done in

$p \times q \times r$ ..

different ways.

Example: Postal delivery

You have just been hired as a Post Delivery person for University of Galway. On your first day, you must travel to seven buildings with letters.

How many different routes are possible?

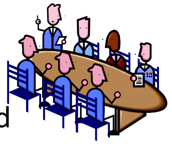
Example: Postal delivery

You have just been hired as a Post Delivery person for University of Galway. On your first day, you must travel to seven buildings with letters.

How many different routes are possible?

7x6x5x4x3x2x1=5040

Example: Committee Problem

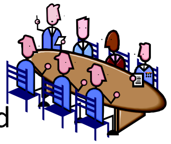


Three members from a 14-member committee are to be randomly selected to serve as chair, vice chair, and secretary.

The first person selected is the chair, the second person selected is to be vice chair, and the third secretary.

How many different committee structures are possible?

Example: Committee Problem



Three members from a 14-member committee are to be randomly selected to serve as chair, vice chair, and secretary.

The first person selected is the chair, the second person selected is to be vice chair, and the third secretary.

How many different committee structures are possible? 14x13x12=2148

Permutations

A **permutation** is an *arrangement* of objects.

We have seen that arranging *n* distinct (different) objects can be done in *n(n-1)(n-2)…3.2.1* different ways.

This calculation is often written using the ***factorial*** symbol. If *n* is an integer, the factorial symbol *n!* is defined as *n! = n(n-1)(n-2)…3.2.1*

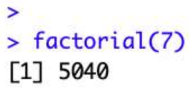
E.g. 3! = 3.2.1 = 6 E.g. 2! = 2.1 = 2

Example: Postal delivery

You have just been hired as a Post Delivery person for University of Galway. On your first day, you must travel to seven buildings with letters.

How many different routes are possible?

7x6x5x4x3x2x1=5040=7!



```
>
> factorial(7)
[1] 5040
```

Permutations

A permutation can also be an arrangement of *r* objects chosen from *n* distinct (different) objects where replacement in the selection is not allowed.

The symbol, P_r^n , represents the number of permutations of *r* objects selected from *n* objects.

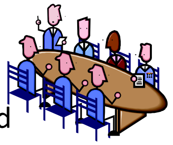
The calculation is given by the formula:

$$P_r^n = \frac{n!}{(n-r)!}$$

Example: Committee Problem

Three members from a 14-member committee are to be randomly selected to serve as chair, vice chair, and secretary.

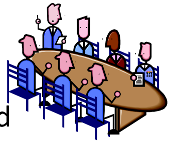
The first person selected is the chair, the second person selected is to be vice chair, and the third secretary.



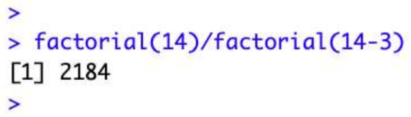
Example: Committee Problem

Three members from a 14-member committee are to be randomly selected to serve as chair, vice chair, and secretary.

The first person selected is the chair, the second person selected is to be vice chair, and the third secretary.



$$P_3^{14} = \frac{14!}{(14-3)!}$$



```
>
> factorial(14)/factorial(14-3)
[1] 2184
>
```

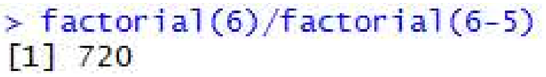
Example: Renmore U12 Soccer

- 6 players available, {A,B,C,D,E,F}
- 5 a side competition
- Number of permutations of 6 players choosing 5 at a time ?

Renmore U12 Soccer

- 6 players available, {A,B,C,D,E,F}
- 5 a side competition
- Number of permutations of 6 players choosing 5 at a time ?

$${}^6P_5 = \frac{n!}{(n-r)!} = \frac{6!}{(6-5)!} = \frac{6!}{1!} = 720$$



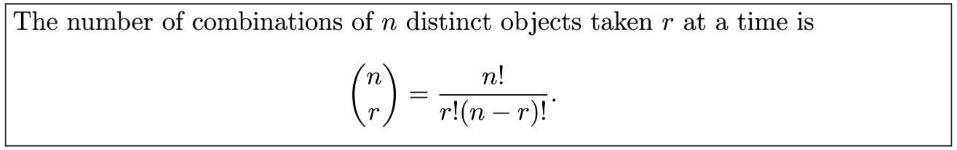
```
> factorial(6)/factorial(6-5)
[1] 720
```

Renmore U12 Soccer

- 6 players available, {A,B,C,D,E,F}
- 5 a side competition
- 720 permutations – be careful as order doesn’t matter here !

Team A,B,C,D,E is the same team as E,D,C,B,A ….. lots of double counting

Combinations (when order doesn't matter!)


The number of combinations of n distinct objects taken r at a time is

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

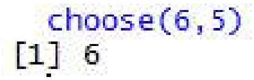
Renmore U12 Soccer

- 6 players available, {A,B,C,D,E,F}
- 5 a side competition
- Number of combinations (as order doesn't matter) ?

Renmore U12 Soccer

- 6 players available, {A,B,C,D,E,F}
- 5 a side competition
- Number of combinations ?

$$^{6}C_{5} = \frac{n!}{r!(n-r)!} = \frac{6!}{5!(6-5)!} = \frac{6!}{5!1!} = 6$$


```
choose(6,5)
[1] 6
```

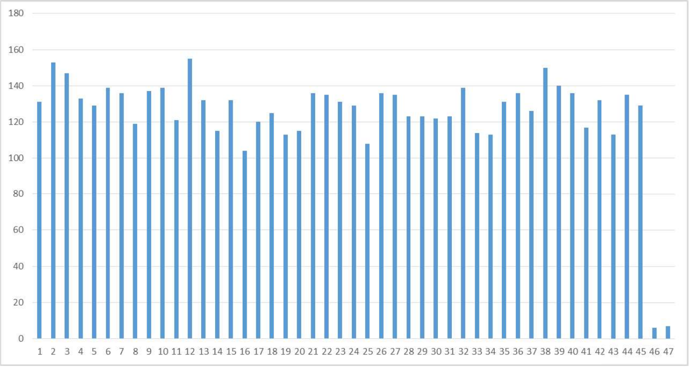
Renmore U12 Soccer

- 6 players available, {A,B,C,D,E,F}
- 5 a side competition
- 6 teams to choose

{A,B,C,D,E}, {A,B,C,D,F}, {A,B,C,E,F}, {A,B,D,E,F}, {A,C,D,E,F}, {B,C,D,E,F}

- Total football, every child gets a chance … what is the probability of any team of this list being the one chosen to start ?

Example: Lottery



Since launch of Lotto 6/45 on November 4th 2006.
Balls 46 and 47 introduced September 3rd 2015.

Knowing when order matters is important …

- https://www.youtube.com/watch?v=wOLxoCF19Ng

Winning the Lotto

- 47 balls available, {1,2,3, … ,47}
- 6 are selected at random
- Number of combinations ?

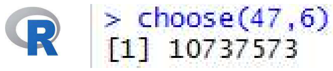
$$ {}^{47}C_6 = \frac{n!}{r!\,(n-r)!} = $$

Winning the Lotto

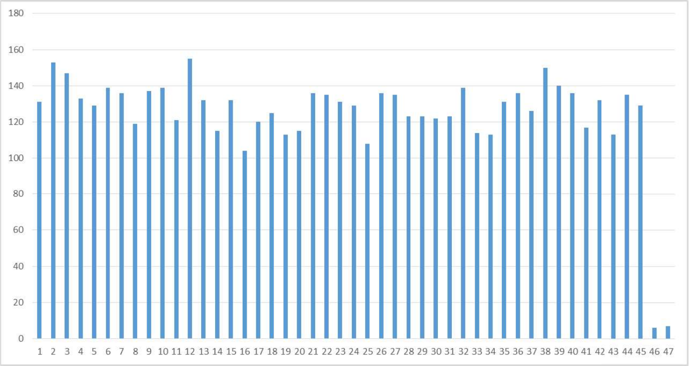
- 47 balls available, {1,2,3, … ,47}
- 6 are selected at random
- Number of combinations ?

$$ {}^{47}C_6 = \frac{n!}{r!(n-r)!} = \frac{47!}{6!(47-6)!} = \frac{47!}{6!41!} $$

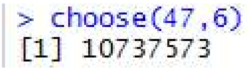
= 10737573



```
> choose(47,6)
[1] 10737573
```



Since launch of Lotto 6/45 on November 4th 2006.
Balls 46 and 47 introduced September 3rd 2015.



```
> choose(47,6)
[1] 10737573
```

For a 4 euro bet, a player fills two lines - two sets of 6 numbers.

What is the probability of winning the Irish Lotto with a 4 euro bet if ordering is not important?

For a 4 euro bet, a player fills two lines - two sets of 6 numbers.

What is the probability of winning the Irish Lotto with a 4 euro bet if ordering is not important?

With two lines there are two chances

$$P(\text{win}) = \frac{2}{10737573} = 0.0000001862618$$

Repetition: *n* non-distinct elements

The number of permutations of n of which
n_1 are of one kind,
n_2 are of a second kind,
...,
and n_k are of a kth kind
is given by

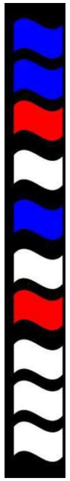
$$\frac{n!}{n_1!n_2!\cdots n_k!}$$

where $n_1 + n_2 + \ldots + n_k = n$.

Repetition example: flags

How many different vertical arrangements are there of 10 flags if 5 are white, 3 are blue and 2 are red?

Solution:

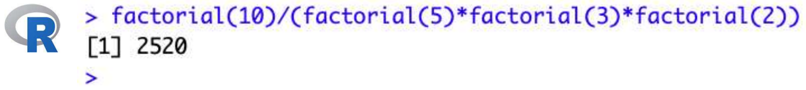


Example: flags

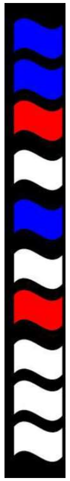
How many different vertical arrangements are there of 10 flags if 5 are white, 3 are blue and 2 are red?

Solution:

$$\frac{10!}{5!\,3!\,2!} = 2520$$



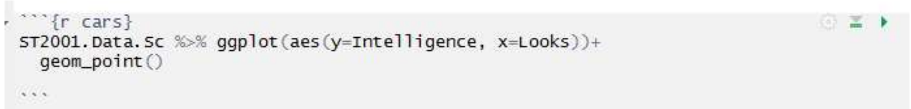
```
> factorial(10)/(factorial(5)*factorial(3)*factorial(2))
[1] 2520
>
```



Joint Events ….

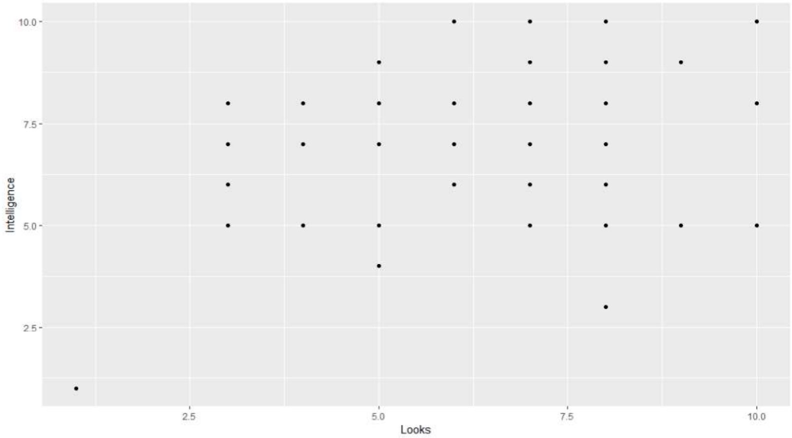
- Class survey …

Simple scatterplot

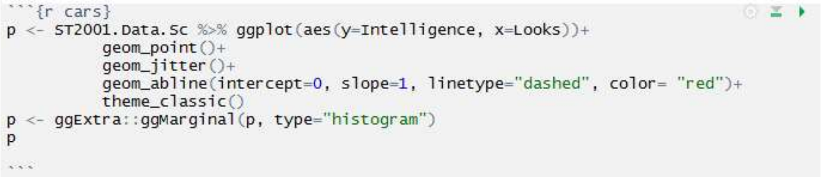


~~~
```{r cars}
ST2001.Data.Sc %>% ggplot(aes(y=Intelligence, x=Looks))+
  geom_point()
```
~~~

Simple scatterplot

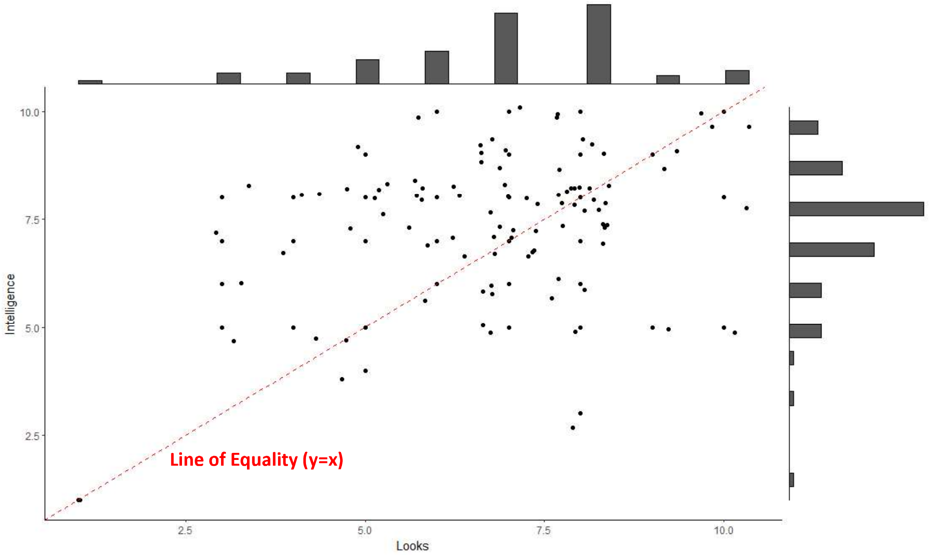


A better scatterplot



~~~
```{r cars}
p <- ST2001.Data.Sc %>% ggplot(aes(y=Intelligence, x=Looks))+
          geom_point()+
          geom_jitter()+
          geom_abline(intercept=0, slope=1, linetype="dashed", color= "red")+
          theme_classic()
p <- ggExtra::ggMarginal(p, type="histogram")
p

```
~~~

Points are jittered (i.e. no longer hidden behind each other) and a line of equality (i.e. y=x) is added as a reference and (marginal) histograms for each variable displayed.



Line of Equality (y=x)

Joint Events ….

- Write down 9 characteristics that your ideal person in life must have.
- Assign a probability to each event.
- Work out the probability of meeting a person with ***all*** characteristics you have listed.

Joint events (and / or)

- Probabilities of joint events can often be determined from the probabilities of the individual events that comprise them.

- Joint events are generated by applying basic set operations to individual events, specifically:
  - Complement of event A is

    $\bar{A}$ = all outcomes ***not*** in A
  - A∪B – ***Union*** of events; A or B or both
  - A∩B – ***Intersection*** of events A and B

  - ***Disjoint*** events cannot occur together, i.e. A∩B = ∅

Example: Rolling a die

- A = score on die is even = { }
- B = score on die is odd = { }
- C = score is greater than 4 = { }
- A∩B =
- A∪B =
- A∩C =
- B∩C =
- (A∩C) ∪ (B∩C) =
- (A∪B) ∩C =

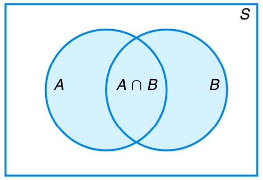
Example: Rolling a die

- A = score on die is even = {2, 4, 6 }
- B = score on die is odd = {1, 3, 5 }
- C = score is greater than 4 = { 5, 6 }
- A∩B = {}= ∅
- A∪B = {1, 2, 3, 4, 5, 6}=S
- A∩C = {6}
- B∩C = {5}
- (A∩C) ∪ (B∩C) = {5,6}
- (A∪B) ∩C = S∩C=C={5,6}

Probability of a Union

- For **any two events** *A* and *B*, the probability of union is given by:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Example ?

Probability of a Union: disjoint events

- For two **disjoint** events **A** and **B**, the probability that one *or* the other occurs is the sum of the probabilities of the two events.

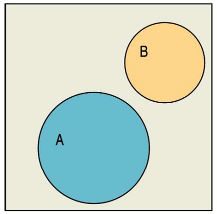
$$P(A \cup B) = P(A) + P(B)$$

provided that **A** and **B** are disjoint.

(also called mutually exclusive)

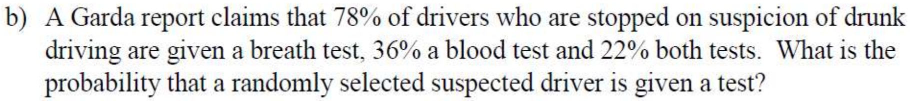
Example ?

Disjoint event share no common outcomes
A ∩ B = Ø



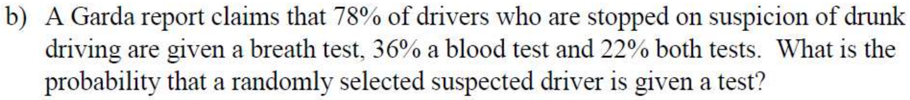
Two disjoint sets, **A** and **B.**

Example …. exam paper 2018

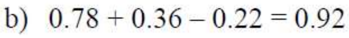


b) A Garda report claims that 78% of drivers who are stopped on suspicion of drunk driving are given a breath test, 36% a blood test and 22% both tests. What is the probability that a randomly selected suspected driver is given a test?

Example …. exam paper 2018



b) A Garda report claims that 78% of drivers who are stopped on suspicion of drunk driving are given a breath test, 36% a blood test and 22% both tests. What is the probability that a randomly selected suspected driver is given a test?



b) $0.78 + 0.36 - 0.22 = 0.92$

Example: Draw a card; event A – an Ace; event B – a heart

	B	$\bar{B}$	Total
A	$\frac{1}{52}$	$\frac{3}{52}$	$\frac{4}{52}$
$\bar{A}$	$\frac{12}{52}$	$\frac{36}{52}$	$\frac{48}{52}$
Total	$\frac{13}{52}$	$\frac{39}{52}$	1.00

Table of joint probabilities

	B	$\bar{B}$	Total
A	$P(A\cap B)$	$P(A\cap \bar{B})$	$P(A)$
$\bar{A}$	$P(\bar{A}\cap B)$	$P(\bar{A}\cap \bar{B})$	$P(\bar{A})$
Total	$P(B)$	$P(\bar{B})$	1.00

Are events disjoint (mutually exclusive) ?

- If P(A U B) is greater than 1 then you know you have made a mistake and the events were not mutually exclusive (i.e. there is an intersection).
- Domain knowledge is needed here …

Intersections (A and B)

Multiplication Rule for independent events:

For two independent events **A** and **B**, the probability that *both* **A** and **B** occur is the product of the probabilities of the two events

$$P(A\cap B)=P(A)\times P(B)$$

provided that **A** and **B** are independent.

This means that occurrence of one event has no impact on the probability of occurrence of the other event.

Example: Electronic components

Two electronic components are selected at random from a production line for inspection. It is known that 90% of the components have no defects.

What is the probability that the two inspected components have no defects?

Calculate probability of intersection of A and B

Let A = 1st component no defect, B=2nd component has no defect

$$P(A \cap B)\ ?$$

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B) \\ &= 0.90 \cdot 0.90 \\ &= 0.81 \end{aligned}$$

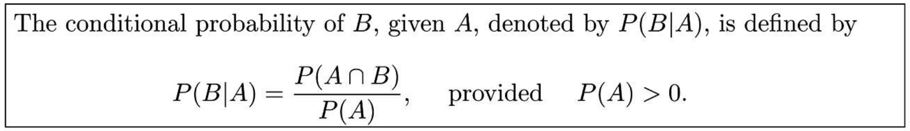
Example: Electronic components with dependence

What if the probability of the second component having no defects changes once we know that the first component had no defects ?

How might this arise ?

Conditional Probability

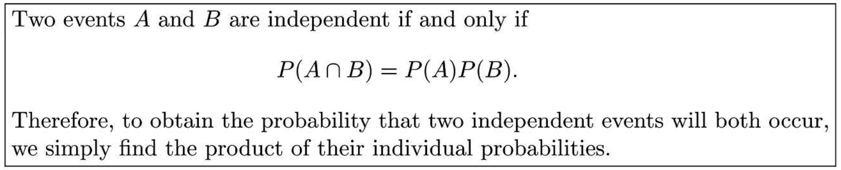
- *P*(*B* | *A*) is the probability of event *B* occurring, given that event *A* has already occurred.



The conditional probability of B, given A, denoted by $P(B|A)$, is defined by

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad \text{provided} \quad P(A) > 0.$$

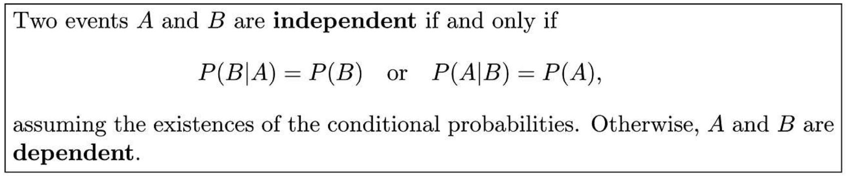
Independence revisited



Two events A and B are independent if and only if

$$P(A \cap B) = P(A)P(B).$$

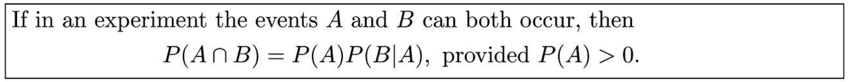
Therefore, to obtain the probability that two independent events will both occur, we simply find the product of their individual probabilities.



Two events A and B are **independent** if and only if

$$P(B|A) = P(B) \quad \text{or} \quad P(A|B) = P(A),$$

assuming the existences of the conditional probabilities. Otherwise, A and B are **dependent**.



If in an experiment the events A and B can both occur, then

$$P(A \cap B) = P(A)P(B|A), \text{ provided } P(A) > 0.$$

Example: Electronic components (dependent events)

Two electronic components are selected at random from a production line for inspection. It is known that the probability that the first component has no defects is 0.90 and that the probability that a second component has no defects given that the first component had no defects is 0.95.

What is the probability that the two inspected components have no defects?

Calculate probability of intersection A and B

Let A = 1st component no defect, B=2nd component has no defect **given** that A had no defect.

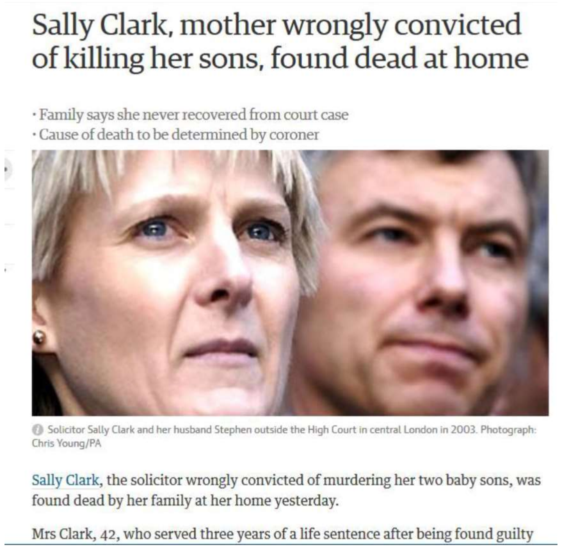
$$P(A \cap B)\ ?$$

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B \mid A) \\ &= 0.90 \times 0.95 \\ &= 0.855 \end{aligned}$$

Knowing that the first was defect free has increased the probability of both being defect free (i.e. from 0.81 to 0.855)

Are events independent ?

- Assuming independence is a HUGE assumption
- The product will be less than 1 so your answer will always make sense but is unlikely to be correct!!



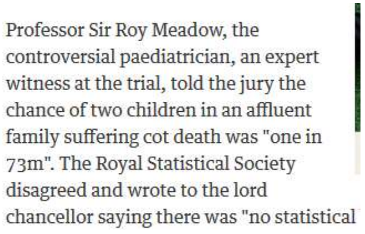
Sally Clark, mother wrongly convicted of killing her sons, found dead at home

- Family says she never recovered from court case
- Cause of death to be determined by coroner

Solicitor Sally Clark and her husband Stephen outside the High Court in central London in 2003. Photograph: Chris Young/PA

Sally Clark, the solicitor wrongly convicted of murdering her two baby sons, was found dead by her family at her home yesterday.

Mrs Clark, 42, who served three years of a life sentence after being found guilty



Professor Sir Roy Meadow, the controversial paediatrician, an expert witness at the trial, told the jury the chance of two children in an affluent family suffering cot death was "one in 73m". The Royal Statistical Society disagreed and wrote to the lord chancellor saying there was "no statistical

Conditional probability when B depends on A

- To find the probability of the event **B** *given* the event **A**, we restrict our attention to the outcomes in **A**. We then find the fraction of *those* outcomes **B** that also occurred.

$$P(\mathbf{B}|\mathbf{A})=\frac{P(\mathbf{A}\cap\mathbf{B})}{P(\mathbf{A})}$$

- Note: *P*(**A**) cannot equal 0, since we know that **A** has occurred.

Example: musical children

At a parents evening at the local Boys school a parent was overheard to say:

"Both of my children are musical"

What is the probability that this parent has two boys?

General Multiplication Rule with dependent events

- The conditional probability can be rewritten to **further** generalise the **multiplication** rule.

$$P(\mathbf{B}|\mathbf{A})=\frac{P(\mathbf{A}\cap\mathbf{B})}{P(\mathbf{A})}$$

Task:

1. **rewrite this formula in terms of P(A ∩ B)**
2. **Use the fact that P(A ∩ B) = P(B ∩ A) and see if you can reverse the conditioning …**

General Multiplication Rule

- The conditional probability can be rewritten to **further** generalise the **multiplication** rule.

$$P(A \cap B) = P(A) \cdot P(B|A)$$
$$P(B \cap A) = P(B)\cdot P(A|B)$$

As P(A ∩ B) = P(B ∩ A) implies

$$P(A).\ P(B|A) = P(B) \cdot P(A|B)$$

Reversing the Conditioning

- This results means that *P*(**A**|**B**) can be calculated once we know *P*(**A**), *P*(**B**, and *P*(**B**|**A**).

- From this information, we can find *P*(**A**|**B**):

$$P(A|B)=\frac{P(B|A)\bullet P(A)}{P(B)} \text{ for } P(B)>0$$

Bayes' Theorem

- Thomas Bayes (1702-1761) was an English mathematician and Presbyterian minister.

- Bayes' theorem states that,

$$P(A|B)=\frac{P(B|A)\bullet P(A)}{P(B)} \text{ for } P(B)>0$$

Recall P(B ∩ A) = P(A ∩ B) implies P(B|A)·P(A) = P(A|B)·P(B)

Diagnostic tests/ Screening

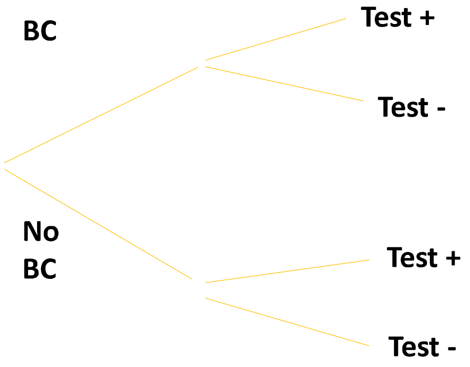
Example: Breast Cancer Screening

- Breast cancer occurs most commonly amongst older women (>60) where it is estimated that 3.65% get breast cancer.
- A mammogram can typically identify correctly 85% of cancer cases (sensitivity) and 95% of cases without cancer (**specificity**).
- If a woman in her 60s gets a positive test what is the probability she has breast cancer?

Breast Cancer Screening tree diagrams

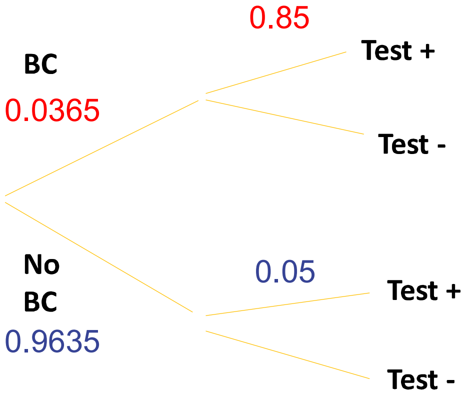
- Think about how many ways can a test come back positive ?
- Tree diagrams are very useful here.

P(BC | +) = ?



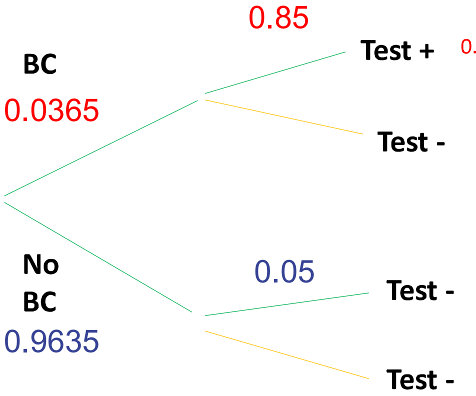
P(Test +) ?

P(BC | +) = ?



P(Test +) ?

P(BC | +) = ?



0.0365 * 0.85 = 0.031

0.9635 * 0.05 = 0.048

P(Test +) ?

$$P(\mathrm{A} \mid \mathrm{B}) = \frac{P(\mathrm{B}|\mathrm{A}).P(\mathrm{A})\cdot}{P(\mathrm{B})} \quad \text{for } P(\mathrm{B}) > 0$$

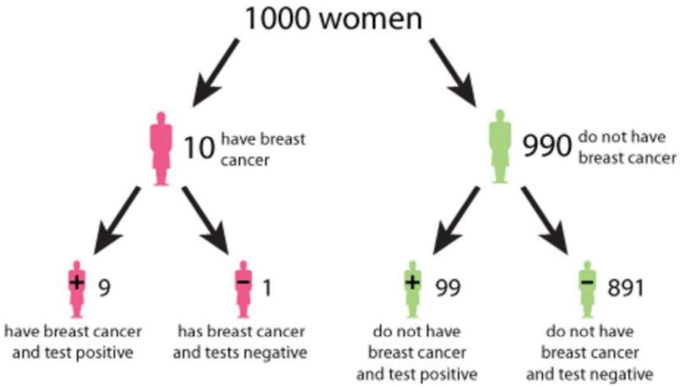
$$P(\mathrm{BC} \mid \mathrm{test}\ +) = \frac{P(\mathrm{test}\ + \mid \mathrm{BC}).P(\mathrm{BC})\cdot}{P(\mathrm{test}\ +)}$$

$$P(\mathrm{BC} \mid \mathrm{test}\ +) = \frac{(0.85\times0.0365)}{(0.85\times0.0365)+(0.05\times0.9635)}$$

$$P(BC \mid test+) = 0.392$$

Interpretation ?

- The probability of having breast cancer given that the test comes back positive is 0.392.
- How would you communicate this result ?


1000 women
10 have breast cancer
990 do not have breast cancer
+ 9
have breast cancer and test positive
- 1
has breast cancer and tests negative
+ 99
do not have breast cancer and test positive
- 891
do not have breast cancer and test negative

Example: printer failures

- A printer manufacturer obtained the following three types of printer failure probabilities:

Hardware P(H) = 0.1,

Software P(S) = 0.6,

Other P(O) = 0.3.

Also, previous experiments suggest

P(F | H) = 0.9,

P(F | S) = 0.2,

P(F | O) = 0.5.

If a failure occurs, determine if it's most likely due to hardware, software, or other.

Types of printer failure

If a failure occurs, determine if it's most likely due to hardware, software, or other.

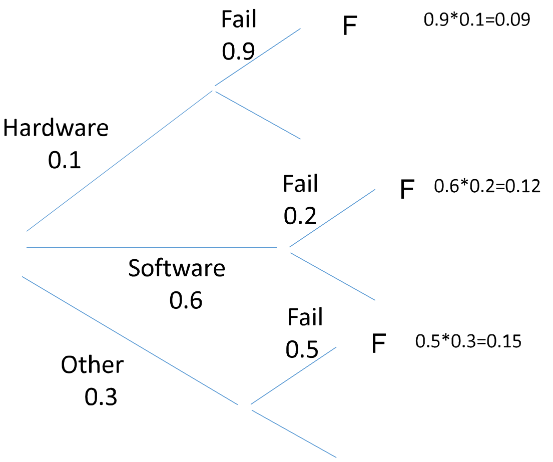
We have P(F | H), P(F | S) and P(F | O)

we need to calculate

P(H | F), P(S | F) and P(O | F)

Start by calculating P(F).

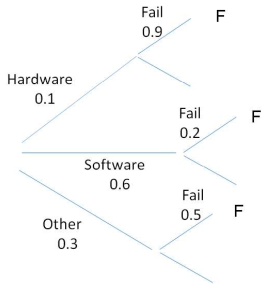
As a tree:



P(F) ?

Calculate probability of failure P(F)

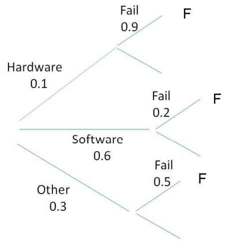
$$P(F) = P(F \mid H)P(H) + P(F \mid S)P(S) + P(F \mid O)P(O)$$
$$= 0.9(0.1) + 0.2(0.6) + 0.5(0.3) = 0.36$$



Now calculate P(H | F) using Bayes Rule

Calculate P(H|F) using Bayes rule

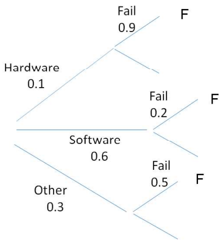
$$P(H \mid F) = \frac{P(F \mid H) \cdot P(H)}{P(F)} = \frac{0.9 \cdot 0.1}{0.36} = 0.250$$



Calculate P(S|F) using Bayes rule

$$P(H\mid F)=\frac{P(F\mid H)\cdot P(H)}{P(F)}=\frac{0.9\cdot 0.1}{0.36}=0.250$$

$$P(S\mid F)=\frac{P(F\mid S)\cdot P(S)}{P(F)}=\frac{0.2\cdot 0.6}{0.36}=0.333$$

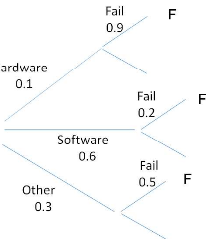


Calculate P(O|F) using Bayes rule

$$P(H\mid F)=\frac{P(F\mid H)\cdot P(H)}{P(F)}=\frac{0.9\cdot 0.1}{0.36}=0.250$$

$$P(S\mid F)=\frac{P(F\mid S)\cdot P(S)}{P(F)}=\frac{0.2\cdot 0.6}{0.36}=0.333$$

$$P(O\mid F)=\frac{P(F\mid O)\cdot P(O)}{P(F)}=\frac{0.5\cdot 0.3}{0.36}=0.417$$

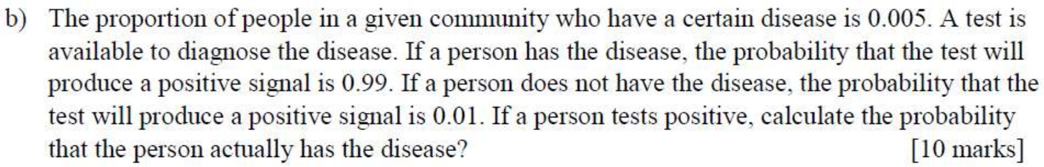


Note that the conditionals given failure add to 1.

Printer failure interpretation

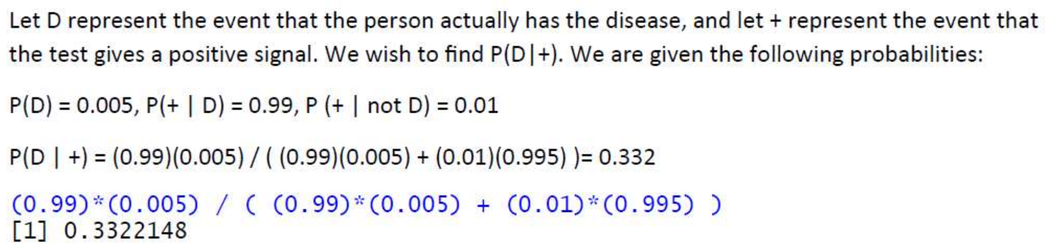
Because *P*(*O* | *F*) is largest, the most likely cause of the problem is in the *other* category.

Screening test for disease: Bayes' Rule example



b) The proportion of people in a given community who have a certain disease is 0.005. A test is available to diagnose the disease. If a person has the disease, the probability that the test will produce a positive signal is 0.99. If a person does not have the disease, the probability that the test will produce a positive signal is 0.01. If a person tests positive, calculate the probability that the person actually has the disease? [10 marks]

Calculate probability of disease given test +



Let D represent the event that the person actually has the disease, and let + represent the event that the test gives a positive signal. We wish to find P(D|+). We are given the following probabilities:

P(D) = 0.005, P(+ | D) = 0.99, P (+ | not D) = 0.01

P(D | +) = (0.99)(0.005) / ((0.99)(0.005) + (0.01)(0.995))= 0.332

```
(0.99)*(0.005) / ( (0.99)*(0.005) + (0.01)*(0.995) )
[1] 0.3322148
```

Bayes Theorem with Total Probability

If E_1, E_2, … E_k are k mutually exclusive and exhaustive events and B is any event,

$$P(E_1 \mid B) = \frac{P(B \mid E_1)P(E_1)}{P(B \mid E_1)P(E_1) + P(B \mid E_2)P(E_2) + \ldots + P(B \mid E_k)P(E_k)}$$

where $P(B) > 0$

Note : Numerator expression is always one of the terms in the sum of the denominator.

Example: Astronauts (Bayes' Rule)

- Astronauts on the shuttle realise that oxygen levels are dropping. There are 3 possible problems that can cause oxygen levels to drop (O): a leak in fuselage (L), malfunctioning oxygen pump (M) and a CO_2 filter in need of replacement (F). It is known that:

  P(L) = 0.02,

  P(M) = 0.49,

  P(F) = 0.49.

Ground crew run simulations to find:

P(O | L) = 1, P(O | M) = 0.4, P(O | F) = 0.6,

What should the astronauts try to fix first ?

Calculate P(L|O)

$$P(L|O) = \frac{P(O|L)P(L)}{P(O|L)P(L)+P(O|M)P(M)+P(O|F)P(F))}$$

$$= \frac{1*0.02}{(1*0.02) + (0.4*0.49) + (0.6*0.49)}$$

$$= \frac{0.02}{0.51}$$

$$= 0.039$$

Calculate P(M|O)

$$P(M|O) = \frac{P(O|M)P(M)}{P(O|L)P(L)+P(O|M)P(M)+P(O|F)P(F))}$$

$$= \frac{0.4*0.49}{0.51}$$

$$= 0.384$$

Calculate P(FIO)

$$P(F|O) = \frac{P(O|F)P(F)}{P(O|L)P(L)+P(O|M)P(M)+P(O|F)P(F))}$$

$$= \frac{0.6*0.49}{0.51}$$

= **0.576**

P(M | O) = 0.384

P(H | O) = 0.039

Astronauts should check the filter first!

Note that P(L|O) + P(M|O) + P(F|O) = 1

Section Summary

- Sample spaces (list by hand or use counting techniques)
  - Permutations and combinations
- Probability
  - Axioms
  - Joint events as Unions (“or”) or intersections (”and”)
  - For unions: mutually exclusive events ?
  - For intersections: independent events ?
  - Conditional probability and Bayes Rule

What Can Go Wrong?

- Beware of probabilities that don’t add up to 1.
  - To be a legitimate probability distribution, the sum of the probabilities for all possible outcomes must total 1.
- Don’t add probabilities of events if they’re not disjoint.
  - Events must be disjoint to use the Addition Rule.

5. Random Variables and Probability Distributions

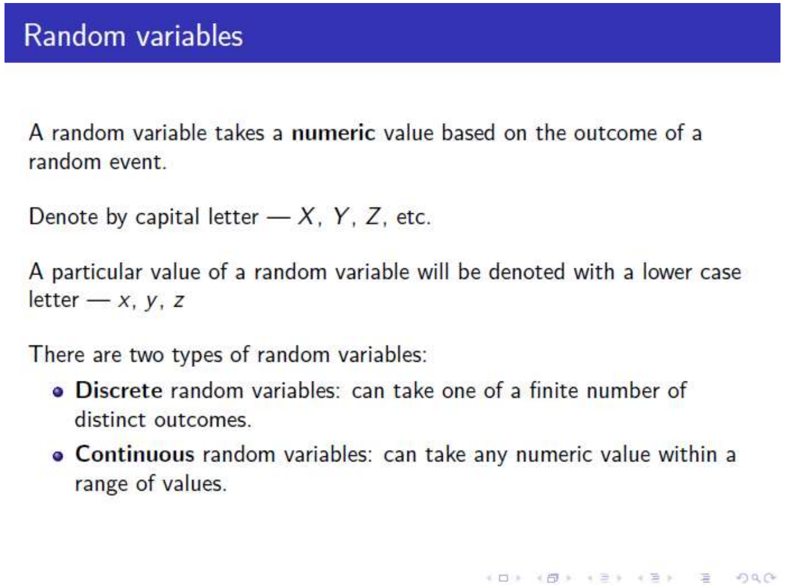
Learning Objectives

1. Determine probabilities from probability mass functions and cumulative distribution functions.
2. Understand the assumptions for probability distributions.
3. Select an appropriate probability distribution to calculate probabilities.
4. Calculate probabilities, means and variances for probability distributions.

Definitions

A **random variable** is a function that associates a real number with each element in the sample space.

The probability distribution of a random variable *X* gives the probability for each value of *X*.



Random variables

A random variable takes a **numeric** value based on the outcome of a random event.

Denote by capital letter — X, Y, Z, etc.

A particular value of a random variable will be denoted with a lower case letter — x, y, z

There are two types of random variables:

- **Discrete** random variables: can take one of a finite number of distinct outcomes.
- **Continuous** random variables: can take any numeric value within a range of values.

Example: Discrete Random Variable

Computer chips may be classed as defective (D) or non-defective (N).
A large batch contains a proportion 0.1 of defectives, and 3 are sampled at random.

Example: Discrete Random Variable

Computer chips may be classed as defective (D) or non-defective (N).
A large batch contains a proportion 0.1 of defectives, and 3 are sampled at random.

The possible outcomes, together with their probabilities are:-

Sample	prob	X
NNN		
DNN		
NDN		
NND		
DDN		
DND		
NDD		
DDD		

Random variable X is the number of defectives in the sample.

Example: Discrete Random Variable

Computer chips may be classed as defective (D) or non-defective (N).
A large batch contains a proportion 0.1 of defectives, and 3 are sampled at random.
The possible outcomes, together with their probabilities are:-

Sample	Prob	X
NNN	$(0.9)^3$	
DNN	$(0.1)(0.9)^2$	
NDN	$(0.9)(0.1))0.9)$	
NND	$(0.9)^2(0.1)$	
DDN	$(0.1)^2(0.9)$	
DND	$(0.1)(0.9)(0.1)$	
NDD	$(0.9)(0.1)^2$	
DDD	$(0.1)^3$	

Random variable X is the number of defectives in the sample.

Example: Discrete Random Variable

Computer chips may be classed as defective (D) or non-defective (N).
A large batch contains a proportion 0.1 of defectives, and 3 are sampled at random.
The possible outcomes, together with their probabilities are:-

Sample	Prob	X
NNN	$(0.9)^3$	0
DNN	$(0.1)(0.9)^2$	1
NDN	$(0.9)(0.1))0.9)$	1
NND	$(0.9)^2(0.1)$	1
DDN	$(0.1)^2(0.9)$	2
DND	$(0.1)(0.9)(0.1)$	2
NDD	$(0.9)(0.1)^2$	2
DDD	$(0.1)^3$	3

Random variable X is the number of defectives in the sample.

Probability model (discrete)

The collection of all possible values of a random variable together with associated probabilities is called the **probability model**

In the example, $\Pr(X = 1)$ can be determined by adding up the probabilities of the 3 sample points associated with the event $X = 1$, etc

x	$\Pr(X = x)$
0	
1	
2	
3	

Probability model (discrete)

The collection of all possible values of a random variable together with associated probabilities is called the **probability model**

In the example, $\Pr(X = 1)$ can be determined by adding up the probabilities of the 3 sample points associated with the event $X = 1$, etc

x	$\Pr(X = x)$
0	$(0.9)^3$
1	$3(0.1)(0.9)^2$
2	$3(0.1)^2(0.9)$
3	$(0.1)^3$

Example

A couple having children will stop when they have a child of each sex or three children.

outcome	GGG	GGB	GB	BG	BBG	BBB
Probability						

Let the random variable X be the number of girls in the family

x	0	1	2	3
$P(X = x)$				

Probability function for a discrete random variable represented pictorially by a **bar graph**.

Example

A couple having children will stop when they have a child of each sex or three children.

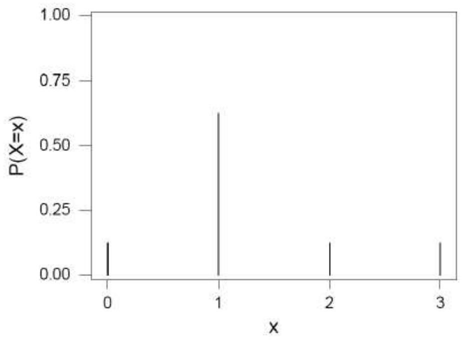
outcome	GGG	GGB	GB	BG	BBG	BBB
Probability	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

Let the random variable X be the number of girls in the family

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{5}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

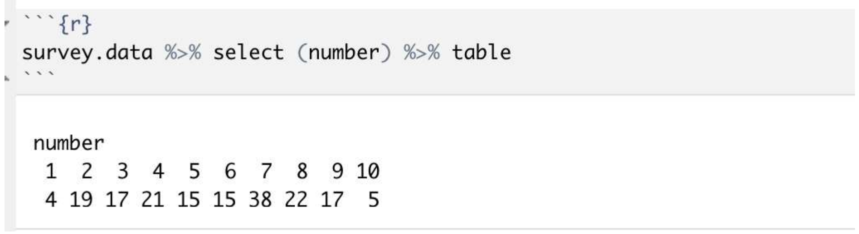
Probability function for a discrete random variable represented pictorially by a **bar graph**.

Probability Function: Bar Graph



Class survey: choosing a number at random

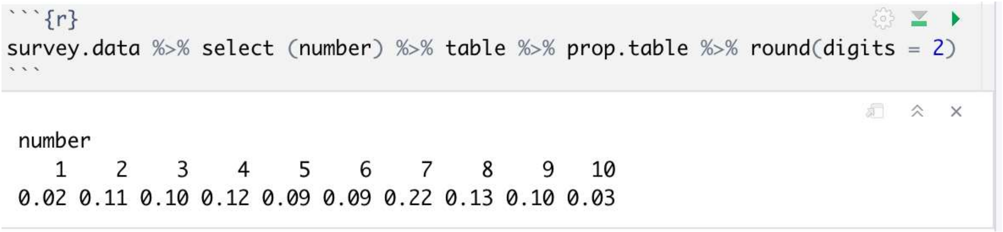
Frequency



~~~
```{r}
survey.data %>% select (number) %>% table
```

number
 1  2  3  4  5  6  7  8  9 10
 4 19 17 21 15 15 38 22 17  5
~~~

Choosing a number at random

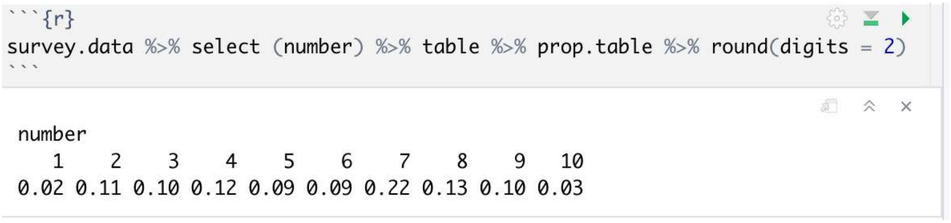
Probability



~~~
```{r}
survey.data %>% select (number) %>% table %>% prop.table %>% round(digits = 2)
```

number
   1    2    3    4    5    6    7    8    9   10
0.02 0.11 0.10 0.12 0.09 0.09 0.22 0.13 0.10 0.03
~~~

Choosing a number at random

Probability



~~~
```{r}
survey.data %>% select (number) %>% table %>% prop.table %>% round(digits = 2)
```

number
   1    2    3    4    5    6    7    8    9   10
0.02 0.11 0.10 0.12 0.09 0.09 0.22 0.13 0.10 0.03
~~~

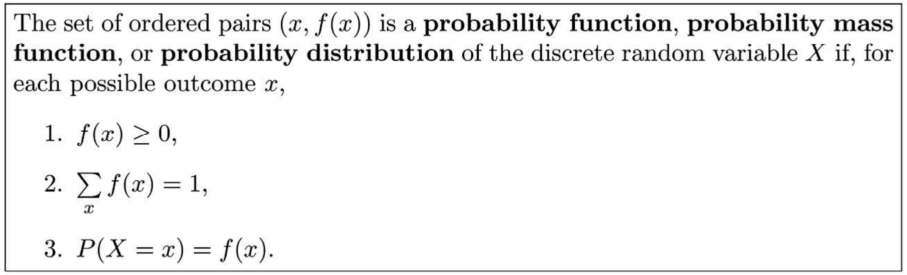
X

***P*(X=x)**

Note that capital X denotes the random variable while small x denotes one of its value

***P*(X=7) = ??**

Discrete Probability Distributions

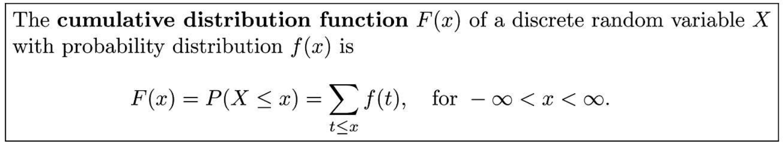


The set of ordered pairs $(x, f(x))$ is a **probability function**, **probability mass function**, or **probability distribution** of the discrete random variable X if, for each possible outcome x,

1. $f(x) \geq 0$,
2. $\sum_x f(x) = 1$,
3. $P(X = x) = f(x)$.

Capital letters for random variables, small letter for one of its values.

Definitions



The **cumulative distribution function** $F(x)$ of a discrete random variable X with probability distribution $f(x)$ is

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \quad \text{for } -\infty < x < \infty.$$

The cumulative distribution function, is the probability that a random variable *X* with a given probability distribution will be found at a value less than or equal to *x*.

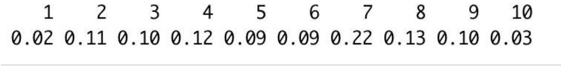
Cumulative Distribution Functions

Consider the probability distribution for the 'choose a number' example. Find the probability of choosing a 3 or less

- The event (*X* ≤ 3) is the total of the events:

  (*X* = 0), (*X* = 1), (*X* = 2), and (*X* = 3).

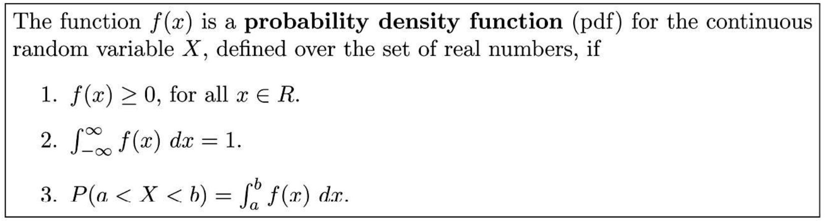
- From the table:



1	2	3	4	5	6	7	8	9	10
0.02	0.11	0.10	0.12	0.09	0.09	0.22	0.13	0.10	0.03

P(*X* ≤ 3) = P(*X* = 0) + P(*X* = 1) + *P*(*X* = 2) + P(*X* = 3) = 0.23

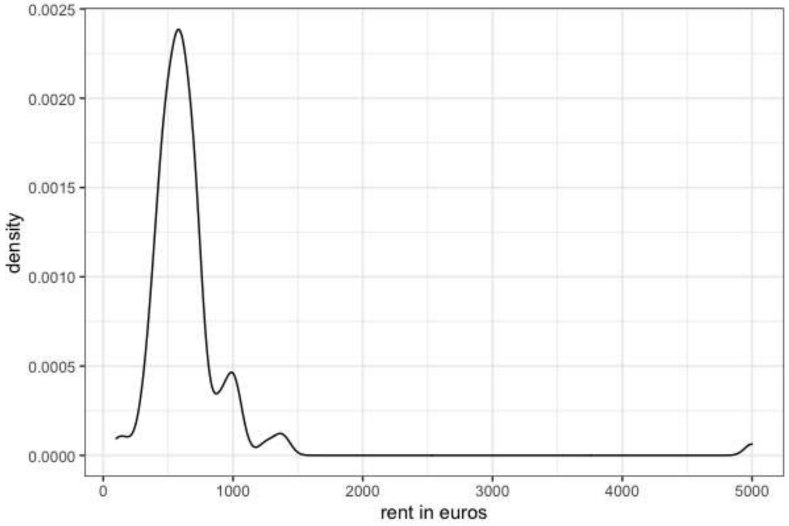
Continuous Probability Distributions



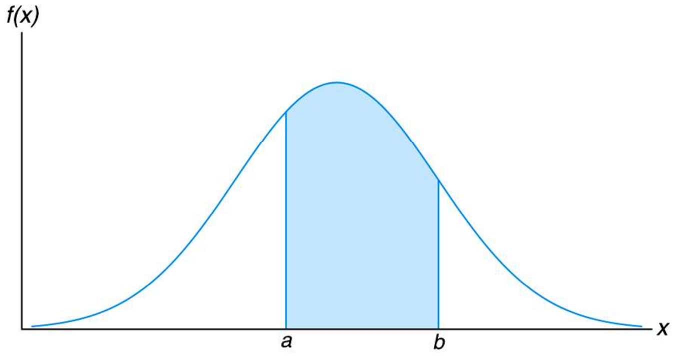
The function $f(x)$ is a **probability density function** (pdf) for the continuous random variable X, defined over the set of real numbers, if

1. $f(x) \geq 0$, for all $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
3. $P(a < X < b) = \int_a^b f(x)\, dx$.

Note P(X = x) = 0 i.e. there is no area exactly at x !

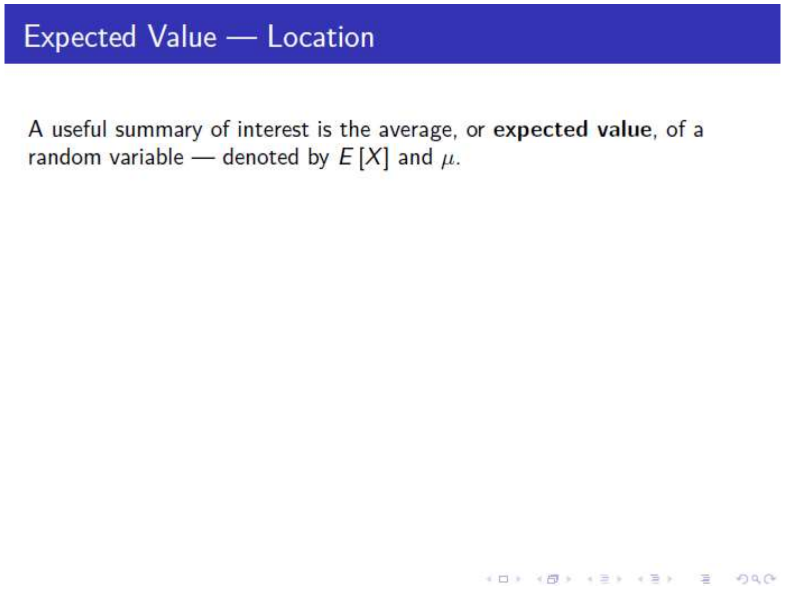


$P(a < X < b)$



Popular discrete and continuous distributions

- Discrete:
  - Binomial
  - Poisson
  - Hypergeometric
- Continuous:
  - Uniform
  - Normal
  - Exponential
- What do they look like ?
- When are they used ?



Expected Value — Location

A useful summary of interest is the average, or **expected value**, of a random variable — denoted by $E[X]$ and μ.

Expected Value — Location

A useful summary of interest is the average, or **expected value**, of a random variable — denoted by $E[X]$ and μ.

The expected value of a random variable can be found by summing the products of each possible value by the probability that it occurs:

$$\mu = E[X] = \sum_x xP(X = x)$$

Expected Value — Location

A useful summary of interest is the average, or **expected value**, of a random variable — denoted by $E[X]$ and μ.

The expected value of a random variable can be found by summing the products of each possible value by the probability that it occurs:

$$\mu = E[X] = \sum_x xP(X = x)$$

Example

:

$$E[No.defectives] = 0\times(0.9)^3+1\times3(0.1)(0.9)^2+2\times3(0.1)^2(0.9)+3\times(0.1)^3 =$$

Variance, Standard Deviation — Spread

The **variance** of a random variable measures the squared deviation from the mean:

$$\sigma^2 = \mathrm{Var}(X) = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 P(X = x)$$

Variance, Standard Deviation — Spread

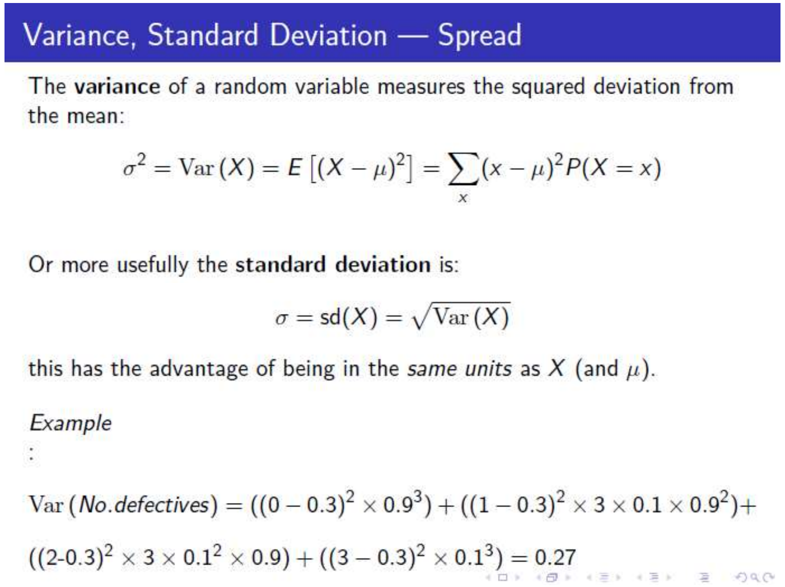
The **variance** of a random variable measures the squared deviation from the mean:

$$\sigma^2 = \mathrm{Var}(X) = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 P(X = x)$$

Or more usefully the **standard deviation** is:

$$\sigma = \mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$$

this has the advantage of being in the *same units* as X (and μ).



Variance, Standard Deviation — Spread

The **variance** of a random variable measures the squared deviation from the mean:

$$\sigma^2 = \mathrm{Var}\,(X) = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 P(X=x)$$

Or more usefully the **standard deviation** is:

$$\sigma = \mathrm{sd}(X) = \sqrt{\mathrm{Var}\,(X)}$$

this has the advantage of being in the *same units* as X (and μ).

Example
:

$\mathrm{Var}\,(No.defectives) = ((0-0.3)^2 \times 0.9^3) + ((1-0.3)^2 \times 3 \times 0.1 \times 0.9^2) +$

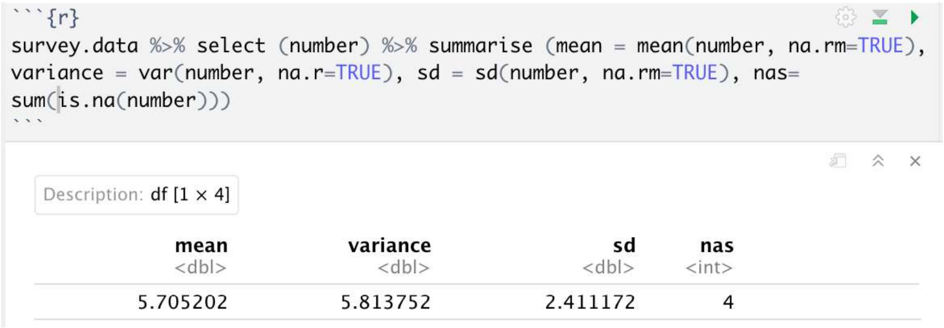
$((2\text{-}0.3)^2 \times 3 \times 0.1^2 \times 0.9) + ((3-0.3)^2 \times 0.1^3) = 0.27$

Variance of a Random Variable

$$\mathrm{Var(X)=E(X^2)\text{-}E^2(X)}$$

Where

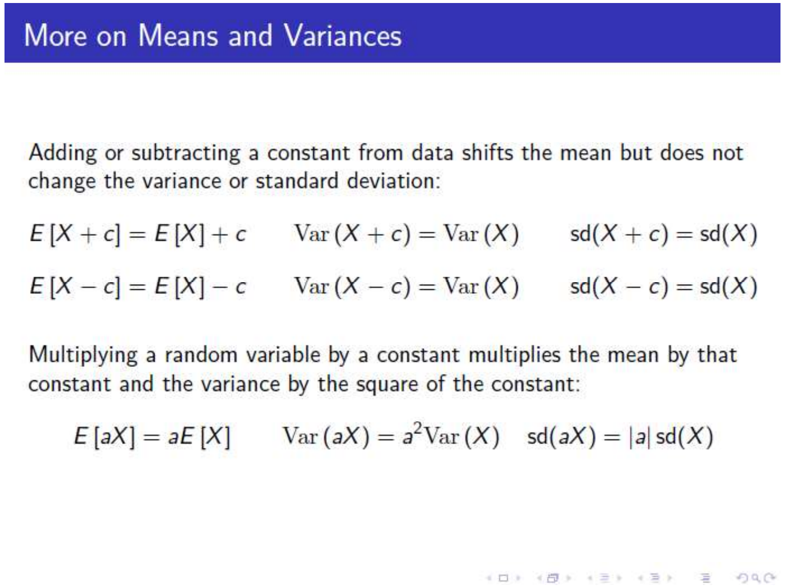
$$E(X^2) = \sum x^2 P(X = x)$$



~~~
```{r}
survey.data %>% select (number) %>% summarise (mean = mean(number, na.rm=TRUE),
variance = var(number, na.r=TRUE), sd = sd(number, na.rm=TRUE), nas=
sum(is.na(number)))
```
~~~

Description: df [1 × 4]

```
      mean     variance          sd    nas
     <dbl>        <dbl>       <dbl>  <int>
  5.705202     5.813752    2.411172      4
```



More on Means and Variances

Adding or subtracting a constant from data shifts the mean but does not change the variance or standard deviation:

$$E[X+c] = E[X]+c \qquad \mathrm{Var}\,(X+c) = \mathrm{Var}\,(X) \qquad \mathrm{sd}(X+c) = \mathrm{sd}(X)$$

$$E[X-c] = E[X]-c \qquad \mathrm{Var}\,(X-c) = \mathrm{Var}\,(X) \qquad \mathrm{sd}(X-c) = \mathrm{sd}(X)$$

Multiplying a random variable by a constant multiplies the mean by that constant and the variance by the square of the constant:

$$E[aX] = aE[X] \qquad \mathrm{Var}\,(aX) = a^2\mathrm{Var}\,(X) \qquad \mathrm{sd}(aX) = |a|\,\mathrm{sd}(X)$$

6. Some Discrete Probability Distributions: the Binomial and Poisson

1

Learning outcomes

- Describe the Binomial distribution and identify when it is applicable
- Calculate Binomial probabilities
- Describe the Poisson distribution and identify when it is applicable
- Calculate Poisson probabilities

2

Links between descriptive stats and probability theory

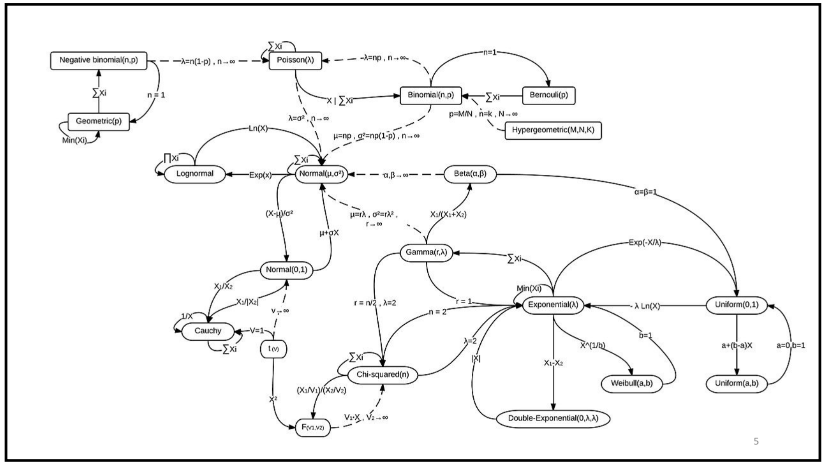
Data	Random variable
x1, x2, … xn	X
Empirical distributions (plots of relative frequencies)	Pmf, pdf
Sample mean	E(X)
Sample variance	Var(X)
Sample sd	Sd(X)

3

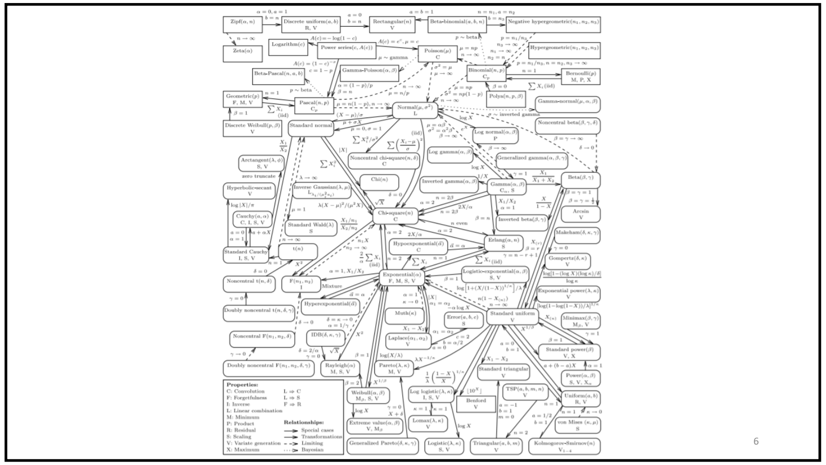
Motivation

- Often, the observations generated by different statistical experiments have the same general type of behaviour.
- In general only a handful of important probability distributions are needed to describe many of the discrete random variables encountered in practice.

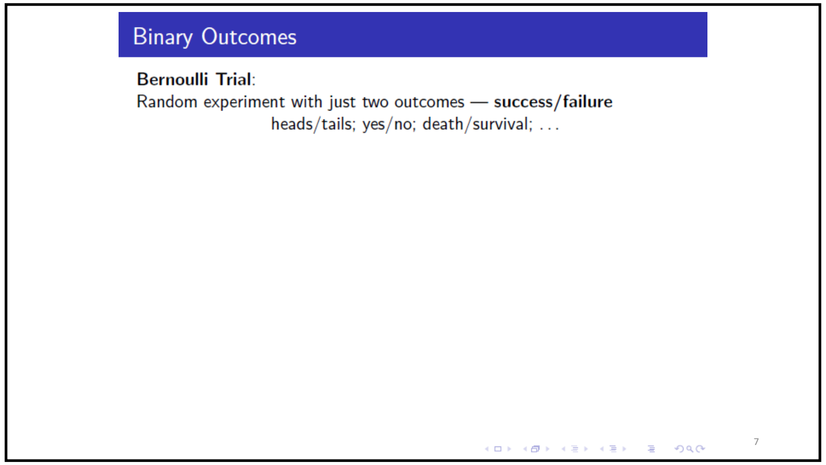
4



5



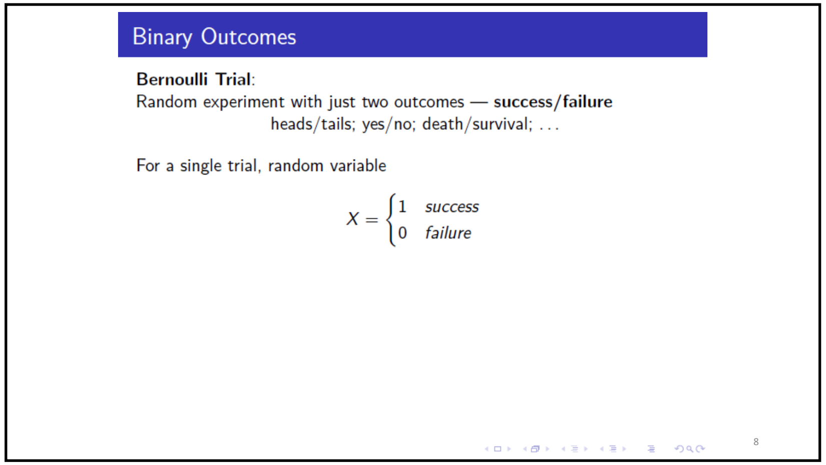
6



Binary Outcomes

Bernoulli Trial:
Random experiment with just two outcomes — **success/failure**
heads/tails; yes/no; death/survival; . . .

7



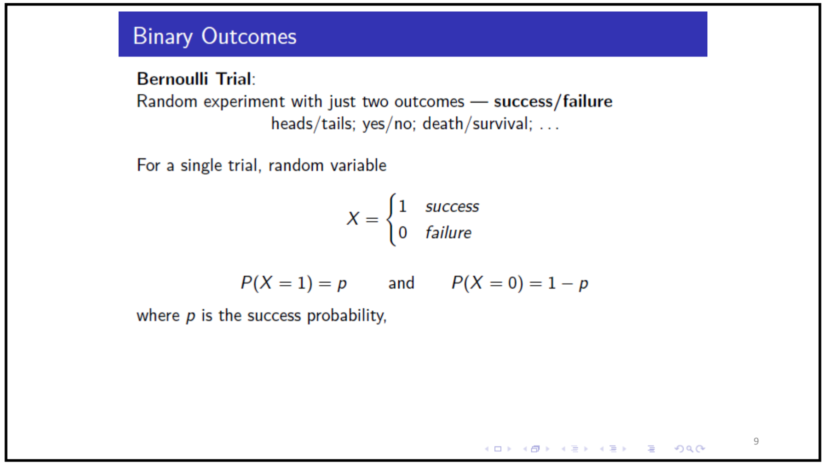
Binary Outcomes

Bernoulli Trial:
Random experiment with just two outcomes — **success/failure**
heads/tails; yes/no; death/survival; . . .

For a single trial, random variable

$$X = \begin{cases} 1 & \textit{success} \\ 0 & \textit{failure} \end{cases}$$

8



Binary Outcomes

Bernoulli Trial:
Random experiment with just two outcomes — **success/failure**
heads/tails; yes/no; death/survival; . . .

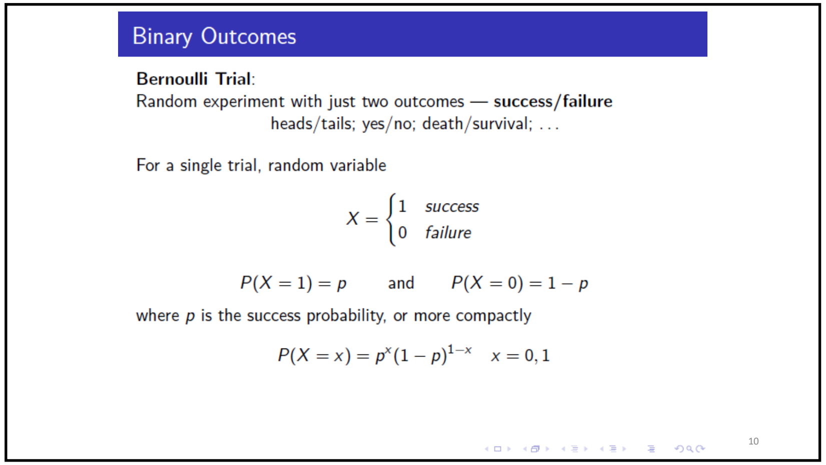
For a single trial, random variable

$$X = \begin{cases} 1 & \textit{success} \\ 0 & \textit{failure} \end{cases}$$

$$P(X = 1) = p \qquad \text{and} \qquad P(X = 0) = 1 - p$$

where p is the success probability,

9



Binary Outcomes

Bernoulli Trial:
Random experiment with just two outcomes — **success/failure**
heads/tails; yes/no; death/survival; . . .

For a single trial, random variable

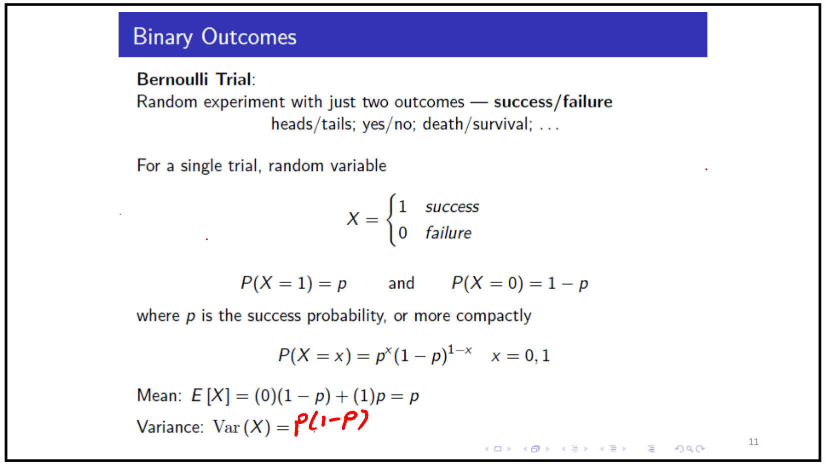
$$X = \begin{cases} 1 & \textit{success} \\ 0 & \textit{failure} \end{cases}$$

$$P(X = 1) = p \qquad \text{and} \qquad P(X = 0) = 1 - p$$

where p is the success probability, or more compactly

$$P(X = x) = p^x(1 - p)^{1-x} \quad x = 0, 1$$

10



Binary Outcomes

Bernoulli Trial:
Random experiment with just two outcomes — **success/failure**
heads/tails; yes/no; death/survival; . . .

For a single trial, random variable

$$X = \begin{cases} 1 & \textit{success} \\ 0 & \textit{failure} \end{cases}$$

$$P(X = 1) = p \qquad \text{and} \qquad P(X = 0) = 1 - p$$

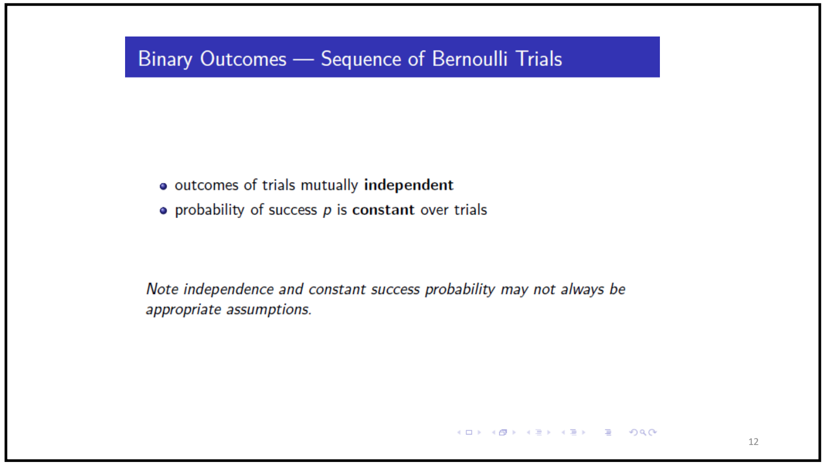
where p is the success probability, or more compactly

$$P(X = x) = p^x(1 - p)^{1-x} \quad x = 0, 1$$

Mean: $E[X] = (0)(1 - p) + (1)p = p$

Variance: $\mathrm{Var}(X) =$ $p(1-p)$

11



Binary Outcomes — Sequence of Bernoulli Trials

- outcomes of trials mutually **independent**
- probability of success p is **constant** over trials

Note independence and constant success probability may not always be appropriate assumptions.

12

Motivating Example: Camera Flash Tests

The time to recharge the flash is tested in **three** mobile phone cameras. The **probability** that a camera passes the test is **0.8**, and the cameras perform independently.

The random variable *X* denotes the number of cameras that pass the test. The last column of the table shows the values of *X* assigned to each outcome of the experiment.

Camera Flash Tests				
Outcome				
Camera #				
1	2	3	Probability	*X*
Pass	Pass	Pass		3
Fail	Pass	Pass		2
Pass	Fail	Pass		2
Fail	Fail	Pass		1
Pass	Pass	Fail		2
Fail	Pass	Fail		1
Pass	Fail	Fail		1
Fail	Fail	Fail		0

What is the probability that the ***first and second cameras pass*** *the test and the third one fails ?*

P(*PPF*) = ?

13

Motivating Example: Camera Flash Tests

What is the probability that the ***first and second cameras pass*** *the test and the third one fails ?*

P(*PPF*) = (0.8)(0.8)(0.2) = 0.128

Each camera test can be treated as a Bernoulli trial.

Probabilities for all other outcomes calculated in a similar fashion.

Camera Flash Tests				
Outcome				
Camera #				
1	2	3	Probability	*X*
Pass	Pass	Pass	0.512	3
Fail	Pass	Pass	0.128	2
Pass	Fail	Pass	0.128	2
Fail	Fail	Pass	0.032	1
Pass	**Pass**	**Fail**	**0.128**	2
Fail	Pass	Fail	0.032	1
Pass	Fail	Fail	0.032	1
Fail	Fail	Fail	0.008	0
			1.000	

What is the probability that two cameras pass the test in three trials ?

14

Motivating Example: Camera Flash Tests

What is the probability that two cameras pass the test in three trials ?

How many **ways** can this event happen ?

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!} = \frac{3!}{2!\,(3-2)!} = \frac{3.2.1}{2.1.1} = 3$$

What is the probability of this event ?

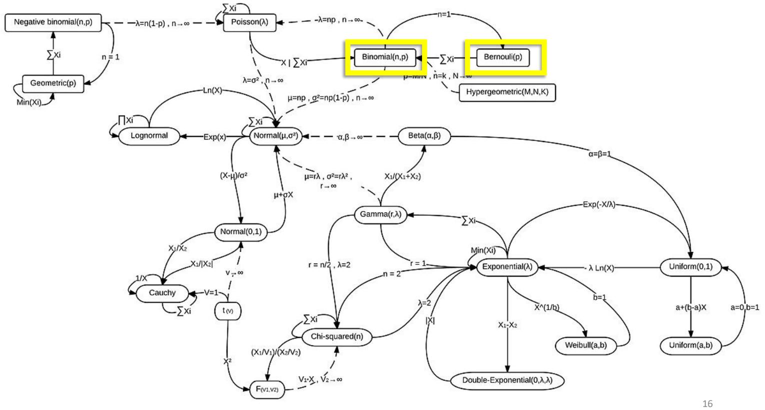
0.128 for **each** of the three ways

probability = 3(0.128) = 0.384

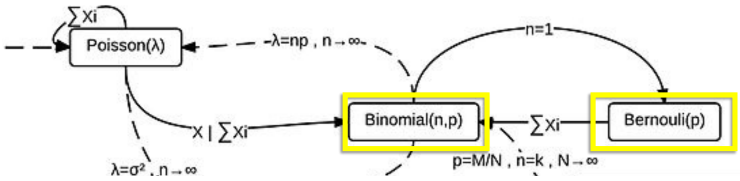
Camera Flash Tests				
Outcome				
Camera #				
1	2	3	Probability	*X*
Pass	Pass	Pass	0.512	3
Fail	Pass	Pass	0.128	2
Pass	Fail	Pass	0.128	2
Fail	Fail	Pass	0.032	1
Pass	Pass	Fail	0.128	2
Fail	Pass	Fail	0.032	1
Pass	Fail	Fail	0.032	1
Fail	Fail	Fail	0.008	0
			1.000	

This is an example of the Binomial Distribution.

15

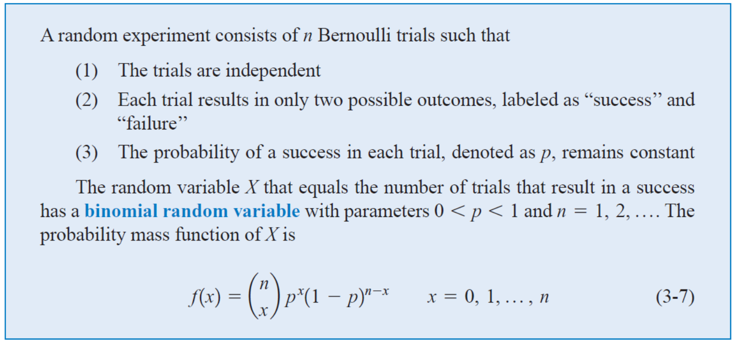


16



17

Binomial Distribution



A random experiment consists of n Bernoulli trials such that

(1) The trials are independent

(2) Each trial results in only two possible outcomes, labeled as "success" and "failure"

(3) The probability of a success in each trial, denoted as p, remains constant

The random variable X that equals the number of trials that result in a success has a **binomial random variable** with parameters $0 < p < 1$ and $n = 1, 2, \ldots$. The probability mass function of X is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots, n \qquad \text{(3-7)}$$

18

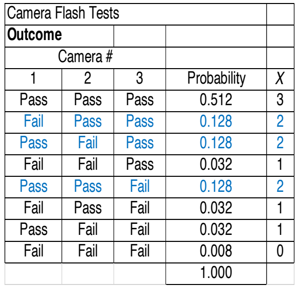
Motivating Example: Camera Flash Tests

Calculate the probability of 2 passes in three tests.

We are given that n = 3 and *p* = 0.8.

Use the Binomial distribution formula where *X* is the number of passes:

$P(X = 2) =$



Camera Flash Tests				
Outcome				
Camera #				
1	2	3	Probability	*X*
Pass	Pass	Pass	0.512	3
Fail	Pass	Pass	0.128	2
Pass	Fail	Pass	0.128	2
Fail	Fail	Pass	0.032	1
Pass	Pass	Fail	0.128	2
Fail	Pass	Fail	0.032	1
Pass	Fail	Fail	0.032	1
Fail	Fail	Fail	0.008	0
			1.000	

19

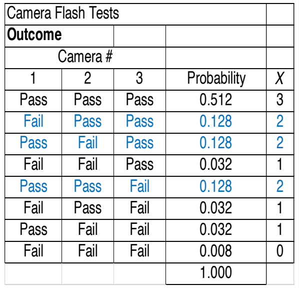
Motivating Example: Camera Flash Tests

Calculate the probability of 2 passes in three tests.

We are given that n = 3 and *p* = 0.8.

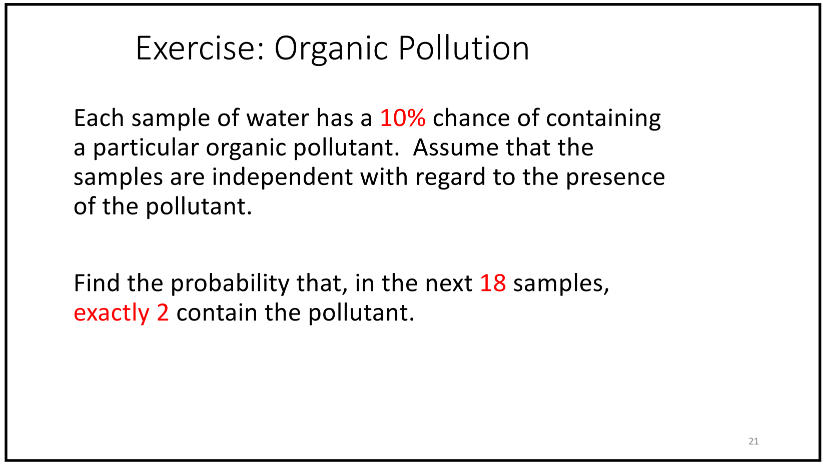
Use the Binomial distribution formula where *X* is the number of passes:

$$\begin{aligned} P(X = 2) &= \binom{3}{2}(0.8)^2(0.2)^1 \\ &= 3(0.128) \\ &= 0.384 \end{aligned}$$



Camera Flash Tests				
Outcome				
Camera #				
1	2	3	Probability	*X*
Pass	Pass	Pass	0.512	3
Fail	Pass	Pass	0.128	2
Pass	Fail	Pass	0.128	2
Fail	Fail	Pass	0.032	1
Pass	Pass	Fail	0.128	2
Fail	Pass	Fail	0.032	1
Pass	Fail	Fail	0.032	1
Fail	Fail	Fail	0.008	0
			1.000	

20



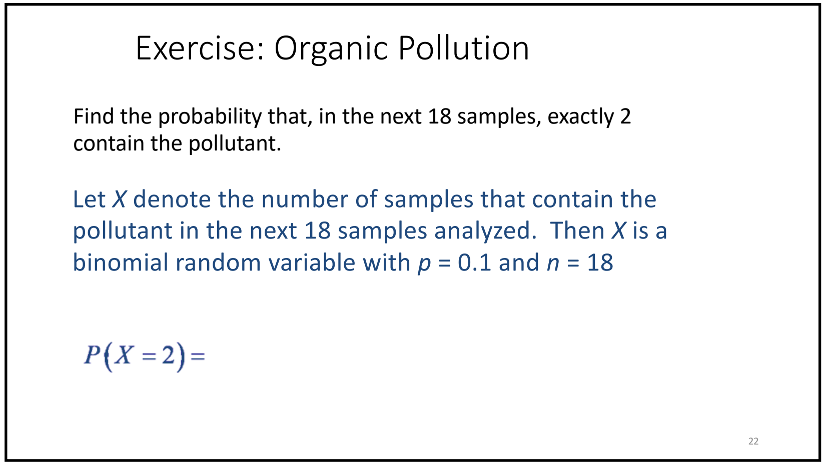
Exercise: Organic Pollution

Each sample of water has a 10% chance of containing a particular organic pollutant. Assume that the samples are independent with regard to the presence of the pollutant.

Find the probability that, in the next 18 samples, exactly 2 contain the pollutant.

21

21



Exercise: Organic Pollution

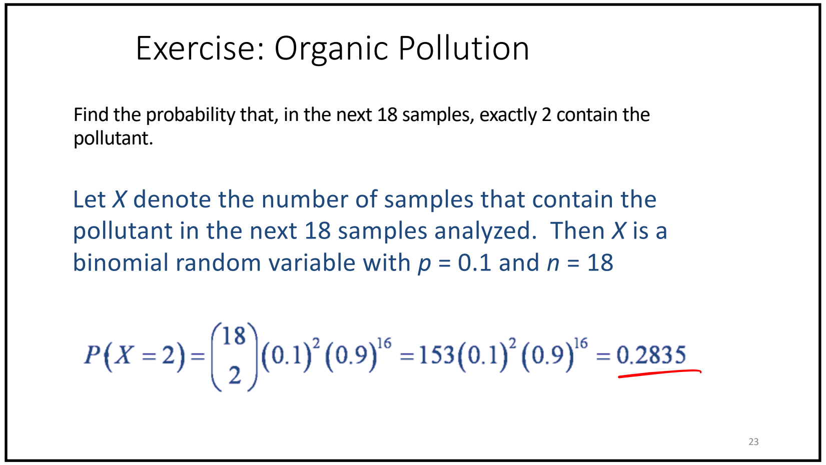
Find the probability that, in the next 18 samples, exactly 2 contain the pollutant.

Let *X* denote the number of samples that contain the pollutant in the next 18 samples analyzed. Then *X* is a binomial random variable with *p* = 0.1 and *n* = 18

$$P(X=2)=$$

22

22



Exercise: Organic Pollution

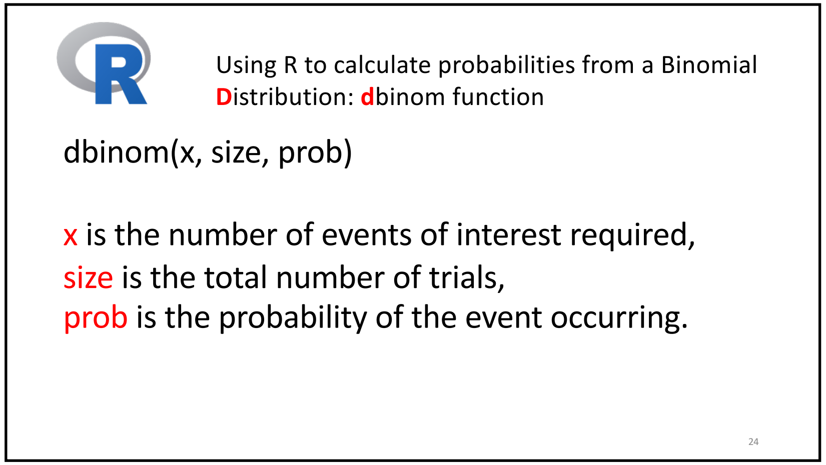
Find the probability that, in the next 18 samples, exactly 2 contain the pollutant.

Let *X* denote the number of samples that contain the pollutant in the next 18 samples analyzed. Then *X* is a binomial random variable with *p* = 0.1 and *n* = 18

$$P(X=2)=\binom{18}{2}(0.1)^2(0.9)^{16}=153(0.1)^2(0.9)^{16}=\underline{0.2835}$$

23

23



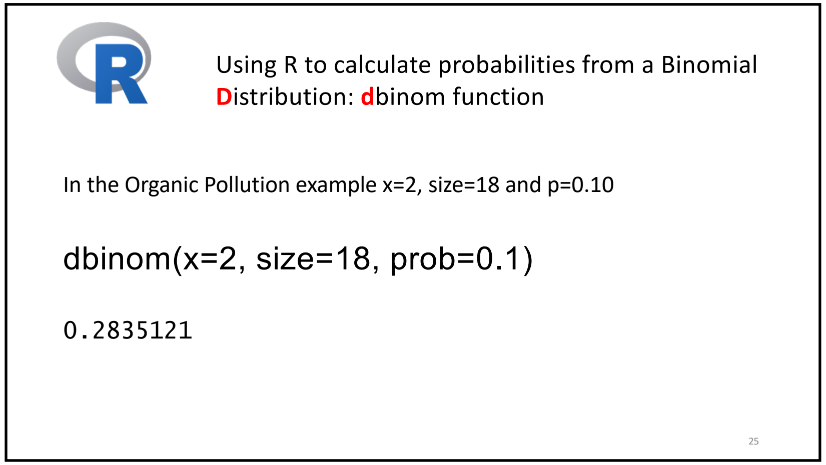
Using R to calculate probabilities from a Binomial Distribution: dbinom function

dbinom(x, size, prob)

x is the number of events of interest required,
size is the total number of trials,
prob is the probability of the event occurring.

24

24

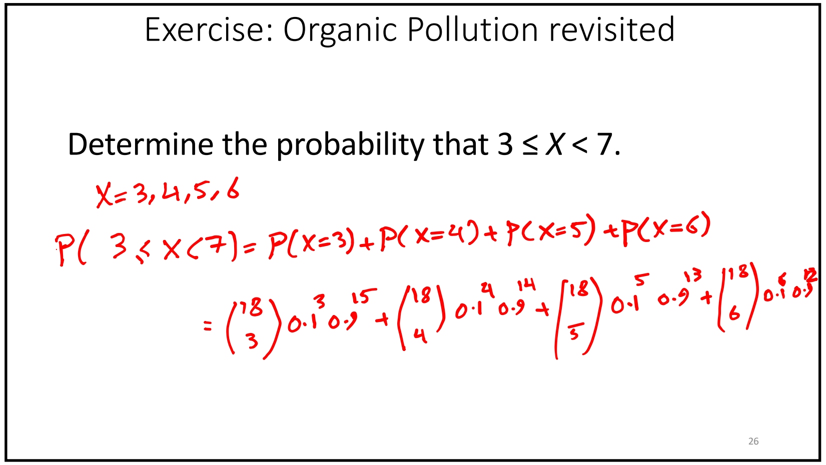


Using R to calculate probabilities from a Binomial Distribution: dbinom function

In the Organic Pollution example x=2, size=18 and p=0.10

```
dbinom(x=2, size=18, prob=0.1)

0.2835121
```

25



Exercise: Organic Pollution revisited

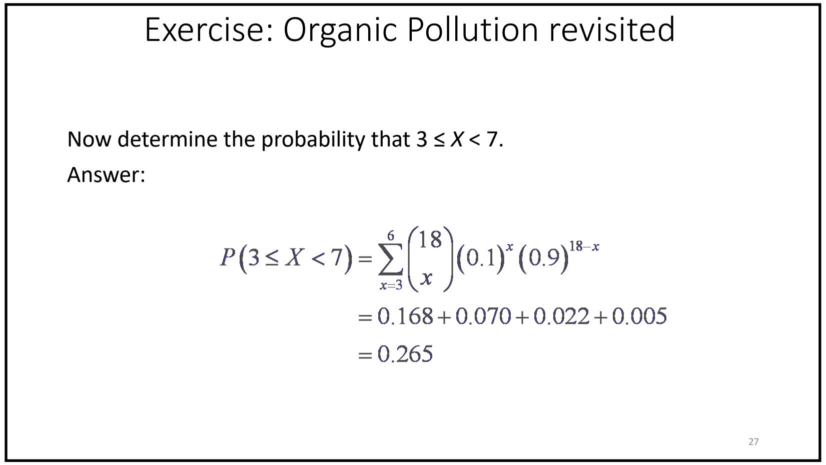
Determine the probability that $3 \le X < 7$.

$X = 3, 4, 5, 6$

$$P(3 \le X < 7) = P(X=3) + P(X=4) + P(X=5) + P(X=6)$$

$$= \binom{18}{3} 0.1^{3} 0.9^{15} + \binom{18}{4} 0.1^{4} 0.9^{14} + \binom{18}{5} 0.1^{5} 0.9^{13} + \binom{18}{6} 0.1^{6} 0.9^{12}$$

26



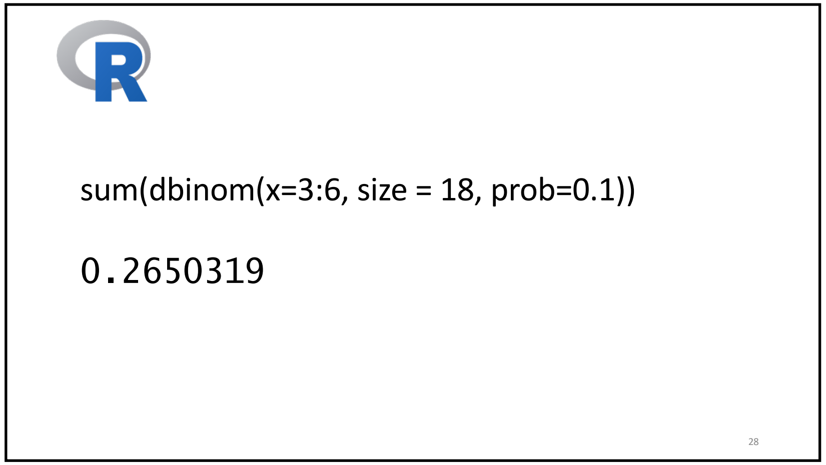
Exercise: Organic Pollution revisited

Now determine the probability that $3 \le X < 7$.

Answer:

$$P(3 \le X < 7) = \sum_{x=3}^{6} \binom{18}{x} (0.1)^{x} (0.9)^{18-x}$$
$$= 0.168 + 0.070 + 0.022 + 0.005$$
$$= 0.265$$

27



```
sum(dbinom(x=3:6, size = 18, prob=0.1))

0.2650319
```

28

Binomial Mean and Variance

If X is a binomial random variable with parameters p and n,

The mean and variance of the binomial distribution $b(x; n, p)$ are

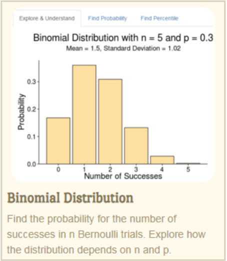
$$\mu = np \text{ and } \sigma^2 = npq.$$

Where q = 1-p.

29

29

Distributions: Explore the Shape, Find Probabilities and Percentiles



Binomial Distribution

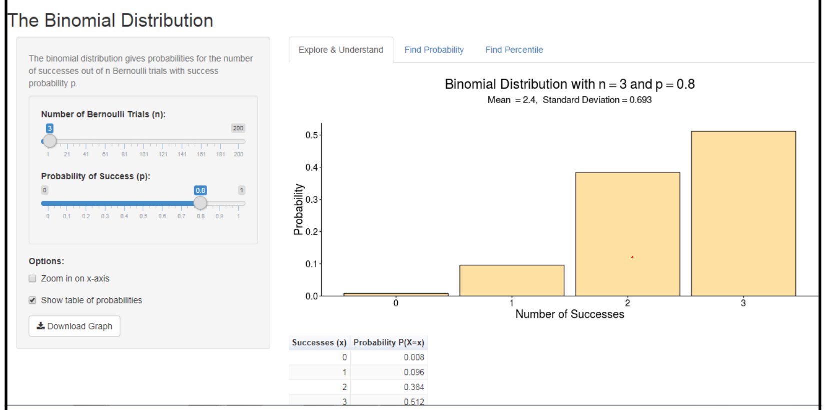
Find the probability for the number of successes in n Bernoulli trials. Explore how the distribution depends on n and p.

http://www.artofstat.com

Use this app to explore different scenarios for a random variable following a Binomial distribution

30

30



The Binomial Distribution

The binomial distribution gives probabilities for the number of successes out of n Bernoulli trials with success probability p.

Number of Bernoulli Trials (n): 3 (range 1–200)

Probability of Success (p): 0.8 (range 0–1)

Options:

☐ Zoom in on x-axis

☑ Show table of probabilities

Binomial Distribution with n = 3 and p = 0.8

Mean = 2.4, Standard Deviation = 0.693

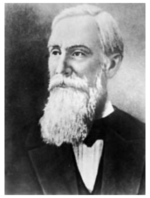
Successes (x)	Probability P(X=x)
0	0.008
1	0.096
2	0.384
3	0.512

31

31

Chebyshev's Inequality

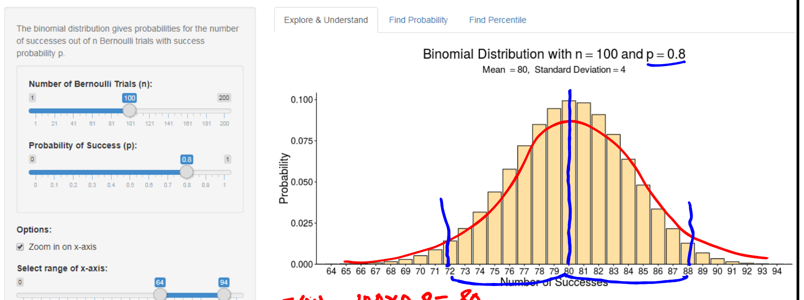
- Chebyshev's inequality provides an estimate as to where a certain % of observations will lie relative to the mean once the **standard deviation** is known.
- For example, at ***least*** 75% of values will lie within two standard deviations of the mean.



32

32

The Binomial Distribution



Explore & Understand | Find Probability | Find Percentile

The binomial distribution gives probabilities for the number of successes out of n Bernoulli trials with success probability p.

Number of Bernoulli Trials (n): 100

Probability of Success (p): 0.8

Options:

☑ Zoom in on x-axis

Select range of x-axis:

64

94

Binomial Distribution with n = 100 and p = 0.8

Mean = 80, Standard Deviation = 4

$E(X) = 100 \times 0.8 = 80$

$Var(X) = 100 \times 0.8 \times 0.2 = 16 \Rightarrow SD(X) = 4$

StatsConsulting.com

- A medical device company needed to calculate the probability that a particular component of their device fails. They have limited bench data which suggests that the probability of failure is 0.15.
- The plan to test 10 devices and want an indication as to the proportion of failures they should expect to see across all devices in the trial.
- What is the number of failures they can *expect* in 10 devices given the probability of failure of a particular device ?

Binomial Mean and Variance

If *X* is a binomial random variable with

parameters *p* and *n*,

The mean and variance of the binomial distribution $b(x; n, p)$ are
$$\mu = np \text{ and } \sigma^2 = npq.$$

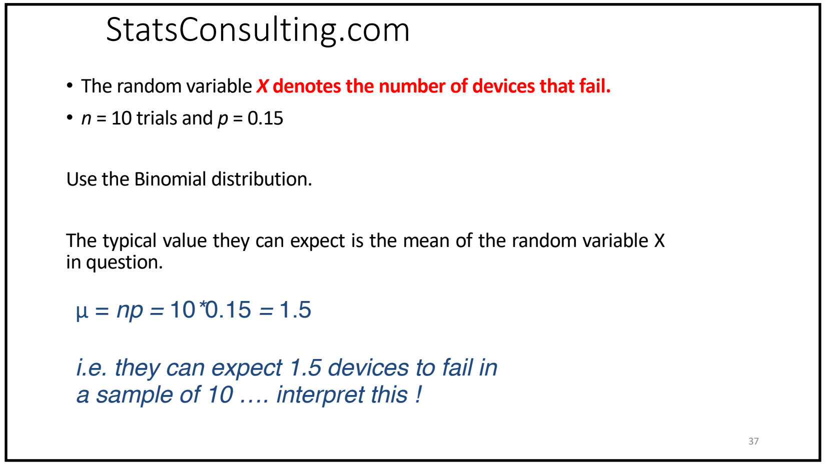
where q=1-p.

StatsConsulting.com

- The random variable ***X* denotes the number of devices that fail.**
- *n* = 10 trials and *p* = 0.15

Use the Binomial distribution.

The typical value they can expect is the mean of the random variable X in question.



StatsConsulting.com

- The random variable ***X* denotes the number of devices that fail.**
- n = 10 trials and p = 0.15

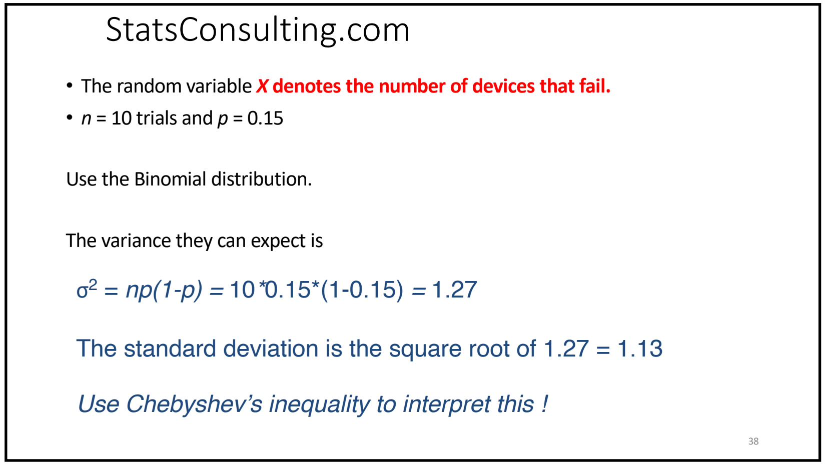
Use the Binomial distribution.

The typical value they can expect is the mean of the random variable X in question.

$\mu = np = 10{*}0.15 = 1.5$

i.e. they can expect 1.5 devices to fail in a sample of 10 …. interpret this !

37



StatsConsulting.com

- The random variable ***X* denotes the number of devices that fail.**
- n = 10 trials and p = 0.15

Use the Binomial distribution.

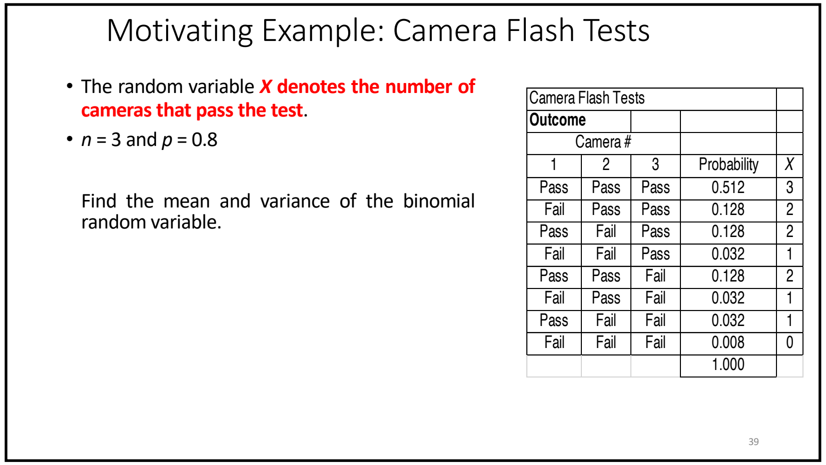
The variance they can expect is

$\sigma^2 = np(1\text{-}p) = 10{*}0.15{*}(1\text{-}0.15) = 1.27$

The standard deviation is the square root of 1.27 = 1.13

Use Chebyshev's inequality to interpret this !

38



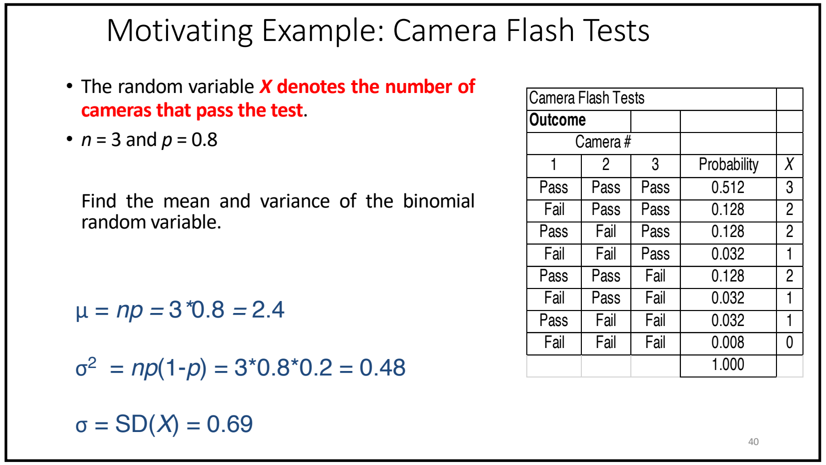
Motivating Example: Camera Flash Tests

- The random variable ***X* denotes the number of cameras that pass the test**.
- n = 3 and p = 0.8

Find the mean and variance of the binomial random variable.

Camera Flash Tests				
Outcome				
Camera #				
1	2	3	Probability	*X*
Pass	Pass	Pass	0.512	3
Fail	Pass	Pass	0.128	2
Pass	Fail	Pass	0.128	2
Fail	Fail	Pass	0.032	1
Pass	Pass	Fail	0.128	2
Fail	Pass	Fail	0.032	1
Pass	Fail	Fail	0.032	1
Fail	Fail	Fail	0.008	0
			1.000	

39



Motivating Example: Camera Flash Tests

- The random variable ***X* denotes the number of cameras that pass the test**.
- n = 3 and p = 0.8

Find the mean and variance of the binomial random variable.

$\mu = np = 3{*}0.8 = 2.4$

$\sigma^2 = np(1\text{-}p) = 3{*}0.8{*}0.2 = 0.48$

$\sigma = \mathrm{SD}(X) = 0.69$

Camera Flash Tests				
Outcome				
Camera #				
1	2	3	Probability	*X*
Pass	Pass	Pass	0.512	3
Fail	Pass	Pass	0.128	2
Pass	Fail	Pass	0.128	2
Fail	Fail	Pass	0.032	1
Pass	Pass	Fail	0.128	2
Fail	Pass	Fail	0.032	1
Pass	Fail	Fail	0.032	1
Fail	Fail	Fail	0.008	0
			1.000	

40

Is the Binomial distribution applicable here ?

Can each trial can be summarized as resulting in either a success or a failure with a fixed probability, assumed independent from trial to trial ?

- A multiple choice test contains 10 questions, each with four choices, and you guess at each question. Let X= the number of questions answered correctly.
- In the next 20 births at a hospital, let X= the number of female births.
- A worn machine tool produces 1% defective parts. Let X=number of defective parts in the next 25 parts produced.
- The probability of ordering a hot chocolate in Mr Waffle is 0.10. A group enters a coffee shop and each member places an order. Let X=number of **hot chocolates** ordered.

Summary so far

- Bernoulli trials and Binomial distribution
- dbinom (in R) and sum(dbinom(start:fininsh, size=, p=) trick
- Mean = np, var=np(1-p)
- When the binomial does and does not apply.
- Oliver's world.

Oliver's world

10,000 products made daily
Probability of a complaint is 0.0001.

What is the probability Oliver will see 10 complaints in a day ?

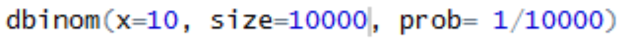
Does the Binomial Distribution apply ?

If you assume it does …. Let X be a random variable representing the number of complaints Oliver will receive in a day.
You are given that n = 10,000 and p=0.0001

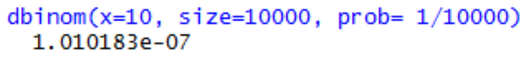
Oliver's world

P(X=10)= $\binom{n}{x}p^x(1-p)^{n-x} = \binom{10,000}{10}0.0001^{10}(1-0.0001)^{10,000-10}$

$$= \ 0.0000001010183$$

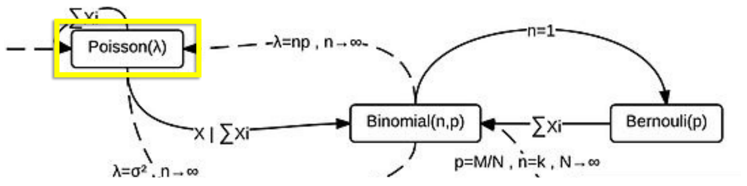


```
dbinom(x=10, size=10000, prob= 1/10000)
```



```
dbinom(x=10, size=10000, prob= 1/10000)
 1.010183e-07
```

Poisson Distribution



Poisson Distribution

- Experiments yielding numerical values of a random variable X, the number of outcomes occurring during a given time interval or in a specified region, are called Poisson experiments.
- The given time interval may be of any length, such as a minute, a day, a week, a month, or even a year.
- A Poisson experiment is derived from the Poisson process and possesses the following properties.

Properties of the Poisson Process

- The number of **outcomes** occurring in one time interval or specified region of space is **independent** of the number that occur in any other disjoint time interval or region. In this sense we say that the Poisson process has no memory.
- The **probability** that a **single outcome** will occur during a very short time interval or in a small region is **proportional** to the **length** of the time interval or the size of the region and does not depend on the number of outcomes occurring outside this time interval or region.
- The probability that **more than one outcome** will occur in such a short time interval or fall in such a small region is **negligible**.

Poisson Distribution

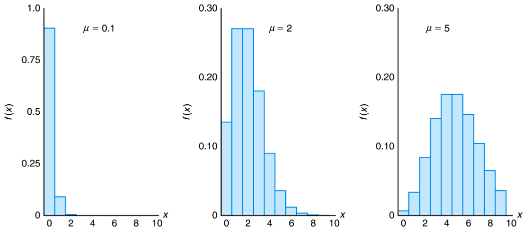
The random variable *X* that equals the **number of events** in a Poisson process is a Poisson random variable with parameter λ > 0, and the probability density function is:

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad \text{for} \quad x = 0,1,2,3,\ldots$$

Mean and Variance of Poisson Distribution

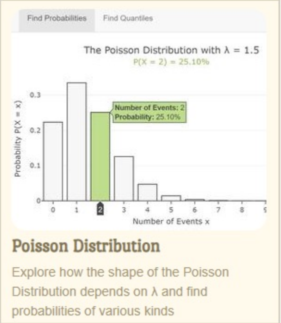
- If λ is the average number of successes occurring in a given time interval or region in the Poisson distribution, then the mean and the variance of the Poisson distribution are both equal to λ.
- Mean = λ, variance = λ
- A one parameter distribution.

Poisson density functions for different means



If the variance is much greater than the mean, then the Poisson distribution would not be a good model for the distribution of the random variable.

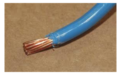
Distributions: Explore the Shape, Find Probabilities and Percentiles



Poisson Distribution

Explore how the shape of the Poisson Distribution depends on λ and find probabilities of various kinds

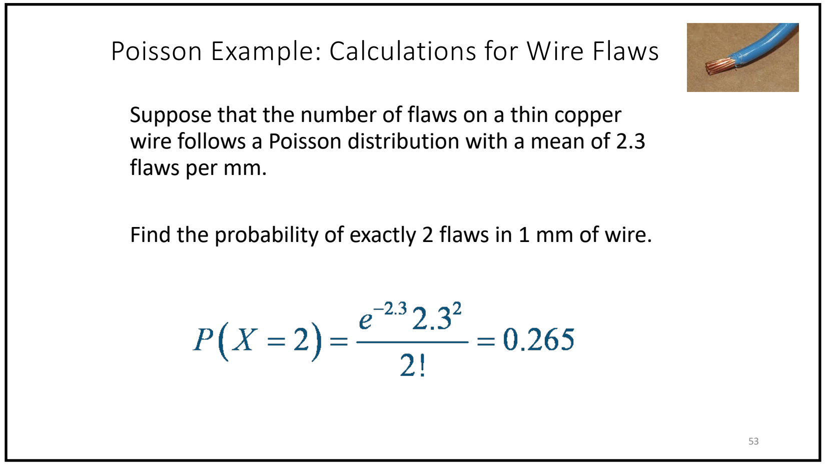
Poisson Example: Calculations for Wire Flaws



Suppose that the number of flaws on a thin copper wire follows a Poisson distribution with a mean of 2.3 flaws per mm.

Find the probability of exactly 2 flaws in 1 mm of wire.

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad \text{for} \quad x = 0,1,2,3,\ldots$$



Poisson Example: Calculations for Wire Flaws

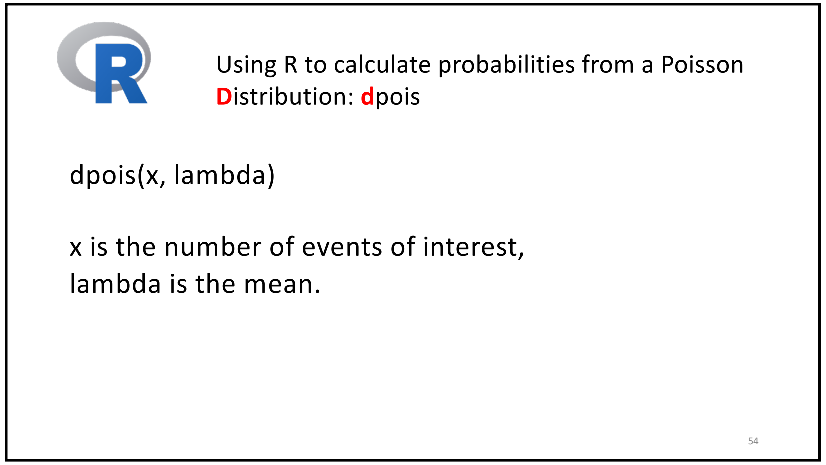
Suppose that the number of flaws on a thin copper wire follows a Poisson distribution with a mean of 2.3 flaws per mm.

Find the probability of exactly 2 flaws in 1 mm of wire.

$$P\left(X=2\right)=\frac{e^{-2.3}2.3^{2}}{2!}=0.265$$

53

53



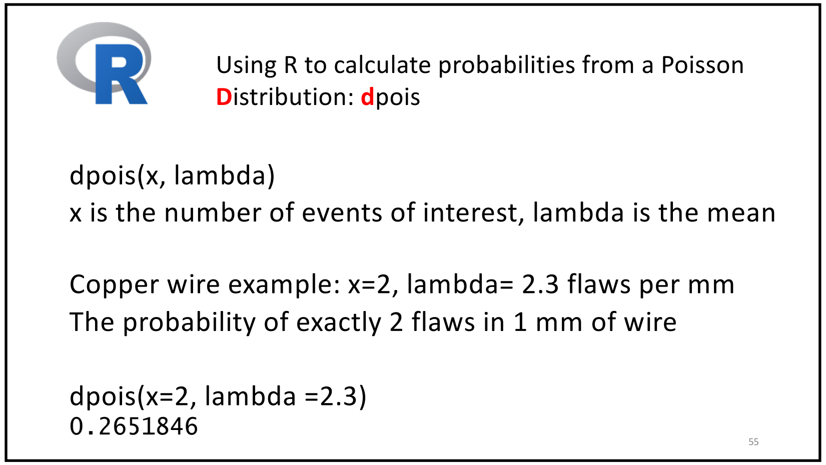
Using R to calculate probabilities from a Poisson **D**istribution: **d**pois

dpois(x, lambda)

x is the number of events of interest,
lambda is the mean.

54

54



Using R to calculate probabilities from a Poisson **D**istribution: **d**pois

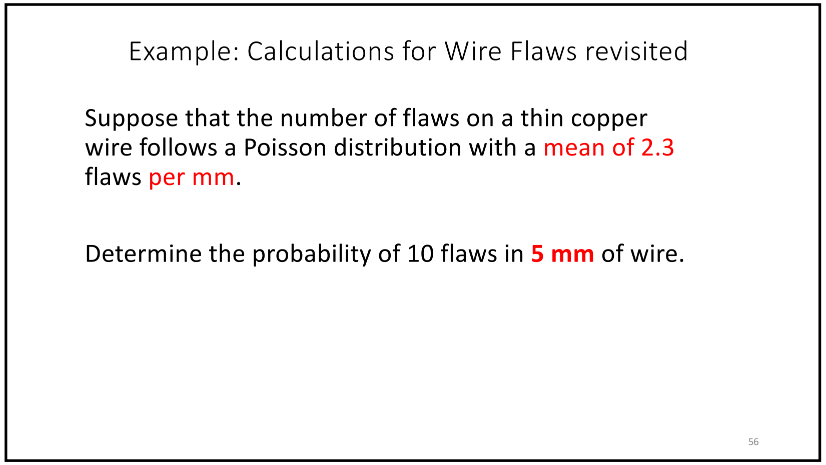
dpois(x, lambda)
x is the number of events of interest, lambda is the mean

Copper wire example: x=2, lambda= 2.3 flaws per mm
The probability of exactly 2 flaws in 1 mm of wire

```
dpois(x=2, lambda =2.3)
0.2651846
```

55

55



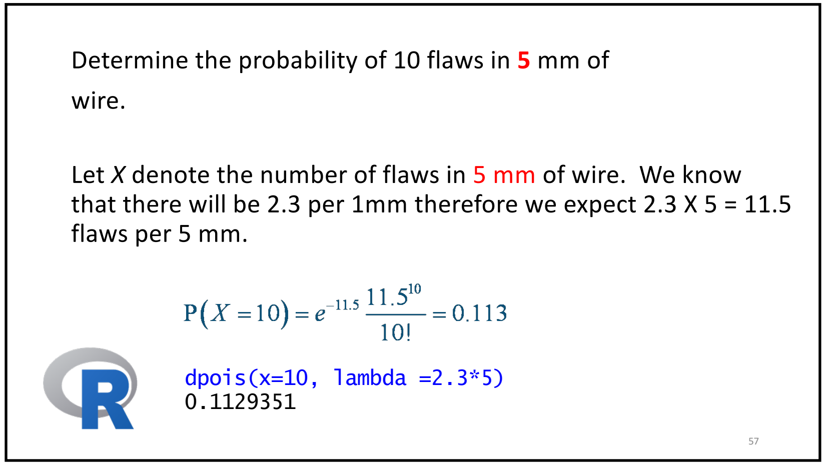
Example: Calculations for Wire Flaws revisited

Suppose that the number of flaws on a thin copper wire follows a Poisson distribution with a mean of 2.3 flaws per mm.

Determine the probability of 10 flaws in **5 mm** of wire.

56

56



Determine the probability of 10 flaws in **5** mm of wire.

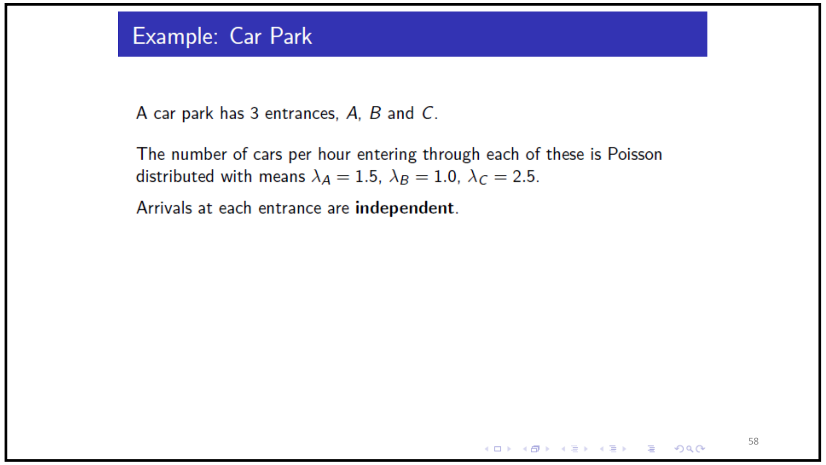
Let X denote the number of flaws in 5 mm of wire. We know that there will be 2.3 per 1mm therefore we expect 2.3 X 5 = 11.5 flaws per 5 mm.

$$\mathrm{P}(X = 10) = e^{-11.5}\frac{11.5^{10}}{10!} = 0.113$$

```
dpois(x=10, lambda =2.3*5)
0.1129351
```

57

57



Example: Car Park

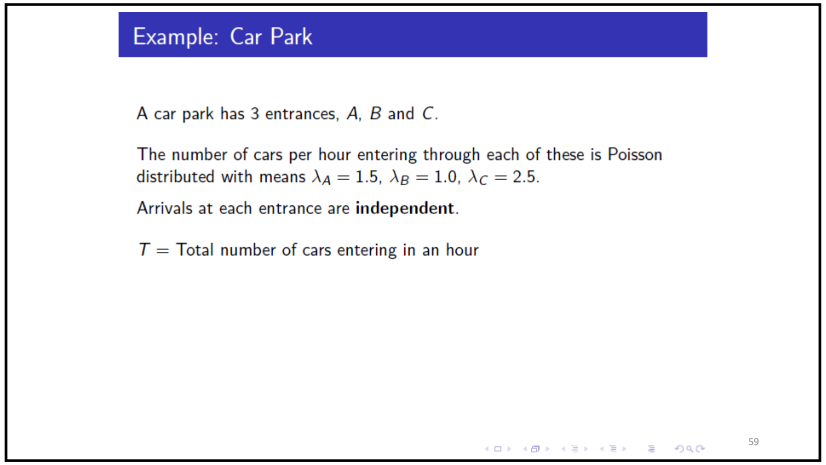
A car park has 3 entrances, A, B and C.

The number of cars per hour entering through each of these is Poisson distributed with means $\lambda_A = 1.5$, $\lambda_B = 1.0$, $\lambda_C = 2.5$.

Arrivals at each entrance are **independent**.

58

58



Example: Car Park

A car park has 3 entrances, A, B and C.

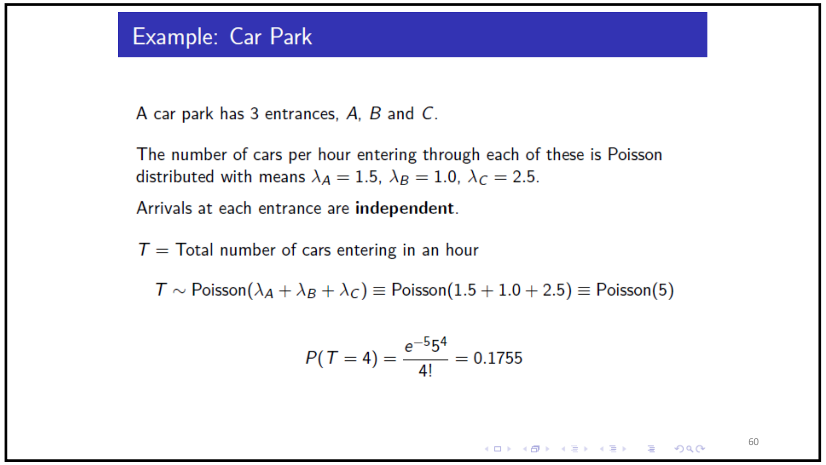
The number of cars per hour entering through each of these is Poisson distributed with means $\lambda_A = 1.5$, $\lambda_B = 1.0$, $\lambda_C = 2.5$.

Arrivals at each entrance are **independent**.

T = Total number of cars entering in an hour

59

59



Example: Car Park

A car park has 3 entrances, A, B and C.

The number of cars per hour entering through each of these is Poisson distributed with means $\lambda_A = 1.5$, $\lambda_B = 1.0$, $\lambda_C = 2.5$.

Arrivals at each entrance are **independent**.

T = Total number of cars entering in an hour

$$T \sim \text{Poisson}(\lambda_A + \lambda_B + \lambda_C) \equiv \text{Poisson}(1.5 + 1.0 + 2.5) \equiv \text{Poisson}(5)$$

$$P(T = 4) = \frac{e^{-5}5^4}{4!} = 0.1755$$

60

60

Sums of Independent Poisson Random Variables

If $X_1, X_2 \ldots, X_n$ are independently Poisson distributed with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$ then

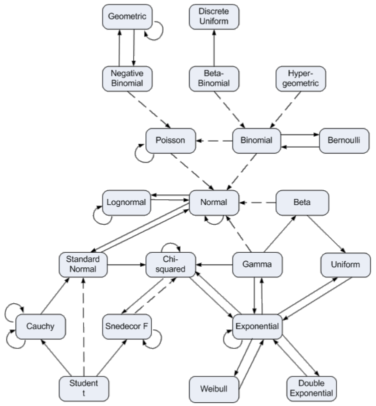
$$T = X_1 + X_2 + \cdots + X_n \quad \text{is} \quad \text{Poisson}(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$$

and

$$E[T] = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

and

$$\text{Var}(T) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$



https://www.johndcook.com/blog/distribution_chart/

The big three ….

- Binomial Distribution
  - In a study involving testing the effectiveness of a new drug, the number of cured patients among all the patients who use the drug approximately follows a binomial distribution
- Geometric Distribution
  - In a statistical quality control problem, the experimenter will signal a shift of the process mean when observational data exceed certain limits. The number of samples required to produce a false alarm follows a geometric distribution.
- Poisson Distribution
  - The number of white cells from a fixed amount of an individual's blood sample is usually random and may be described by a Poisson distribution.

7. The Normal Distribution

Learning Objectives

- Describe features of the Normal distribution
- Describe the effects of changing values of the mean and standard deviation on the normal distribution
- Describe the Empirical Rule and its relationship with the normal distribution
- Describe features of the Standard Normal distribution
- Calculate normal probabilities using z-scores
- Calculate values of a normal random variable given the probability, (using the z-tables in reverse)
- Use R to calculate normal probabilities.

Continuous Probability Distributions Recap

The function $f(x)$ is a **probability density function** (pdf) for the continuous random variable X, defined over the set of real numbers, if

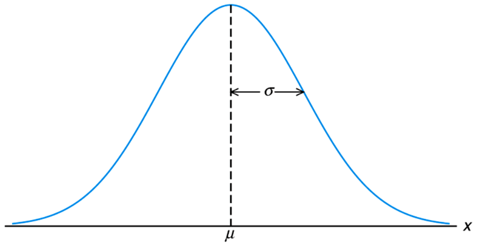
1. $f(x) \geq 0$, for all $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)\ dx = 1$.
3. $P(a < X < b) = \int_a^b f(x)\ dx$.

Note P(X=x) = 0 i.e. there is no area exactly at x !

Normal Distribution

- Also called the Gaussian distribution
- pdf is a bell-shaped curve
- The distribution of many types of observations can be approximated by a Normal – eg consider the relative frequency histograms of
  - Heights
  - Weight
  - IQ, …, etc
- Single mode
- Symmetric
- Model for continuous measurements

The normal distribution



Normal Distribution

A random variable *X* with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

is a normal random variable with parameters μ and σ

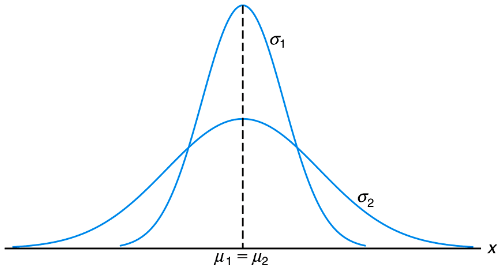
(where - ∞ < μ < ∞ and σ > 0)

Mean

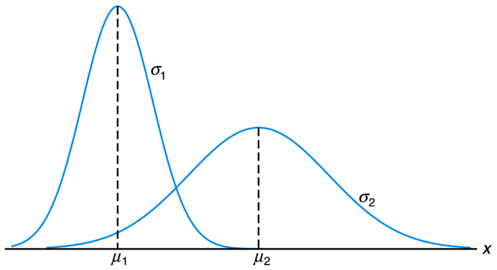
Standard deviation

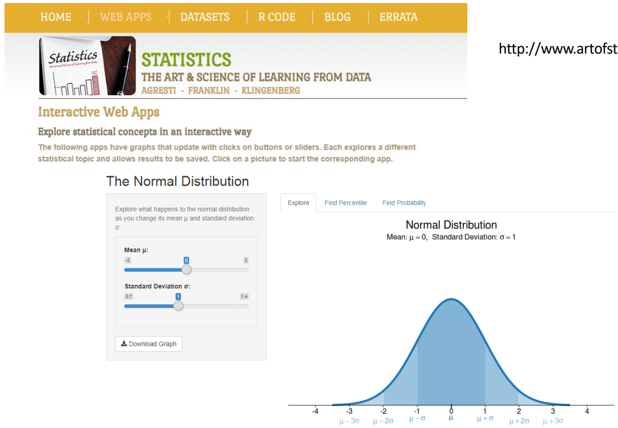
Write X ~ N(μ , σ^2)

Normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$





STATISTICS

THE ART & SCIENCE OF LEARNING FROM DATA

AGRESTI · FRANKLIN · KLINGENBERG

Interactive Web Apps

Explore statistical concepts in an interactive way

The following apps have graphs that update with clicks on buttons or sliders. Each explores a different statistical topic and allows results to be saved. Click on a picture to start the corresponding app.

The Normal Distribution

Explore what happens to the normal distribution as you change its mean μ and standard deviation σ.

Normal Distribution

Mean: μ = 0, Standard Deviation: σ = 1

http://www.artofstat.com

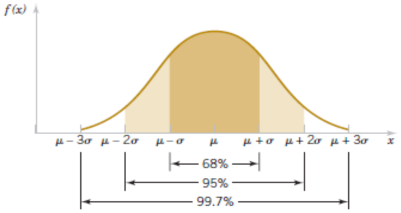
Empirical Rule for a Normal Distribution

For any normal random variable,

$P(\mu - \sigma < X < \mu + \sigma) = 0.6827$

$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$

$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$

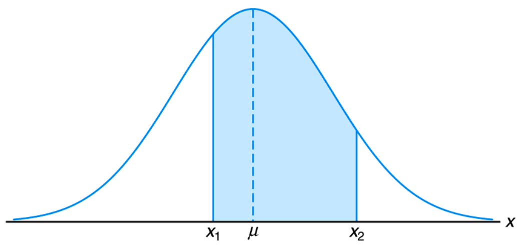


Probabilities associated with a normal distribution

The 68-95-99.7 Rule

- Normal models give us an idea of how extreme a value is by telling us how likely it is to find one that far from the mean.
- It turns out that in a Normal model:
  - about 68% of the values fall within one standard deviation of the mean;
  - about 95% of the values fall within two standard deviations of the mean; and,
  - about 99.7% (almost all!) of the values fall within three standard deviations of the mean.

$P(x_1 < X < x_2)$ = area of the shaded region



Areas under the Normal Curve

- Finding an area under a normal distribution in order to calculate probabilities

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Standardised Z scores.

To convert a random variable X which follows a N(μ , σ^2) to a random variable Z that follows a standard Normal N(0, 1) calculate Z as

$$Z = \frac{X - \mu}{\sigma}$$

Convert X ~ N(100 , 100) to a random variable Z such that Z ~ N(0 , 1)

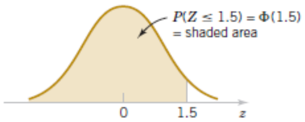
Z scores

- A *z*-score reports the number of standard deviations away from the mean.

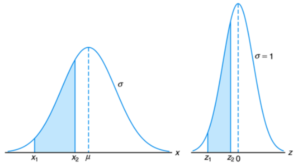
- For example, a Z-score of 2 indicates that the observation is two standard deviations above the mean.

$\Phi(z) = P(Z \le z)$

The cumulative distribution function of a standard normal random variable is denoted as $\Phi(z) = P(Z \le z)$



z	0.00	0.01	0.02	0.03
0	0.50000	0.50399	0.50398	0.51197
⋮		⋮		
1.5	0.93319	0.93448	0.93574	0.93699



Calculating Probabilities for N(0,1)

- Left tail – P(Z < 1.8)
  - Directly from table
- Right tail – P(Z > 1.8)
  - By subtraction P(Z>1.8)=1 – P(Z ≤ 1.8)
- Interval Probabilities – P(1 < Z < 1.8)
  - By difference: P(1 < Z < 1.8)=P(Z<1.8)-P(Z<1)

Normal Probabilities by Hand

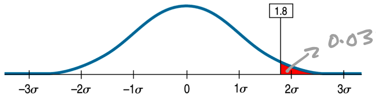
$X \sim N(\mu, \sigma^2)$

- Use a table of the Standard Normal Distribution
- Convert to *z*-scores before using the table. $\rightarrow Z$

$$P(X < k) = P\left(\underbrace{\frac{X-\mu}{\sigma}}_{Z} < \frac{k-\mu}{\sigma}\right) = P\left(Z < \frac{k-\mu}{\sigma}\right)$$

- X ~ N(500,100^2) ; P(X > 680) = P(Z > 1.8)=1-P(Z<1.8)=1-0.9641=0.0359

$$P(X > 680) = P\left(\frac{X-\mu}{\sigma} > \frac{680-\mu}{\sigma}\right)$$
$$= P\left(Z > \frac{680-500}{100}\right)$$
$$= P(Z > 1.8)$$

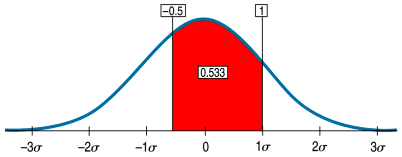


0.0359

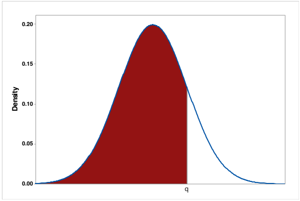
z	.00	.01
⋮	⋮	⋮
1.7	.9554	.9564
~~1.8~~ →	.9641	.9649
1.9	.9713	.9719
⋮	⋮	⋮

Normal: P(-0.5 < Z < 1)=P(Z<1)-P(Z<-0.5)

=0.8413-0.3085=0.5328



Using R to calculate probabilities from a Normal Distribution



pnorm(q=?? , mean= ?? , sd= ??)

pnorm returns the integral from -∞ to q for the pdf of the normal distribution with mean μ and standard deviation σ.

Example: Normal Distribution

Suppose that the current measurements in a strip of wire are assumed to follow a normal distribution with μ = 10 and σ = 2 mA, what is the probability that the current measurement is less than or equal to 9 mA?

$Z = \frac{X-\mu}{\sigma}$

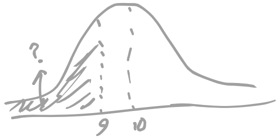
$X \sim N(10, 2^2)$

Plot:

$P(X<9) = P\left(\frac{X-\mu}{\sigma} < \frac{9-\mu}{\sigma}\right)$

$= P\left(Z < \frac{9-10}{2}\right)$

$= P(Z < -0.5) = 0.3085$

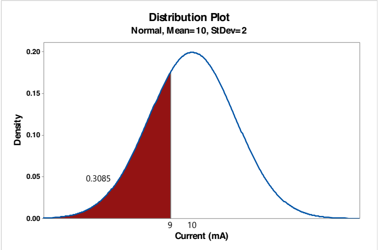


pnorm(9, mean=10, sd=2)=

Example: Normal Distribution

Suppose that the current measurements in a strip of wire are assumed to follow a normal distribution with μ = 10 and σ = 2 mA, what is the probability that the current measurement is less than or equal to 9 mA?

Plot:



Example: Normal Distribution

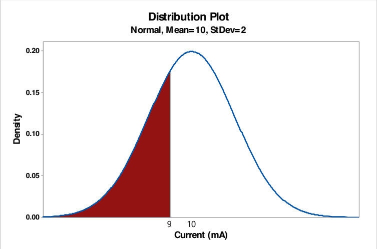
Suppose that the current measurements in a strip of wire are assumed to follow a normal distribution with μ = 10 and σ = 2 mA, what is the probability that the current measurement is less than or equal to 9 mA?

Area:

$$P(-\infty < X \leq 9)$$

$$= \int_{-\infty}^{9} \frac{1}{\sqrt{2\pi 2^2}} exp^{\frac{(x-10)^2}{2(2^2)}} dx$$

$$= P\left(\frac{X-10}{2} < \frac{9-10}{2}\right) = P(Z < -0.5)$$

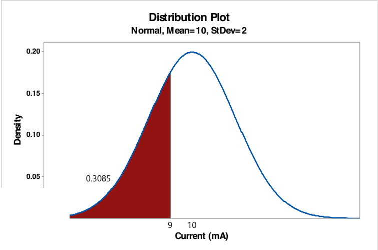


Example: Normal Distribution

Suppose that the current measurements in a strip of wire are assumed to follow a normal distribution with μ = 10 and σ = 2 mA, what is the probability that the current measurement is less than or equal to 9 mA?

Area:

```
> pnorm(q=9, mean=10, sd=2, lower.tail = TRUE)
[1] 0.3085375
> pnorm(-0.5, mean=0, sd=1, lower.tail = T)
[1] 0.3085375
> pnorm(-0.5)
[1] 0.3085375
```



Using R to calculate probabilities from a Normal Distribution

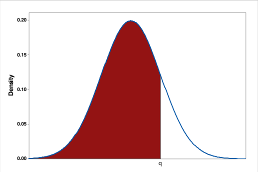
pnorm

pnorm(q=?? , mean= ?? , sd= ??, lower.tail = ??)

pnorm returns the integral from -∞ to q for the pdf of the normal distribution with mean μ and standard deviation σ.

Note: the default is a standardised normal. It means

pnorm(q=??)=pnorm(q=?? , mean= 0, sd= 1, lower.tail = ??)

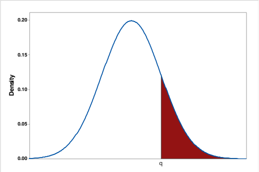


TRUE is the default

pnorm(q=?? , mean= 0, sd= 1, lower.tail = TRUE)

Defaults

Which equals to: pnorm(q=??)

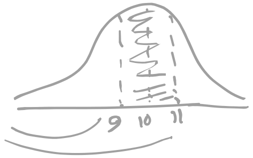


pnorm(q=?? , mean= 0, sd= 1, lower.tail = FALSE)

Example: Normal Distribution

Suppose that the current measurements in a strip of wire are assumed to follow a normal distribution with μ = 10 and σ = 2 mA, what is the probability that the current measurement is between 9 and 11 mA?

Plot:

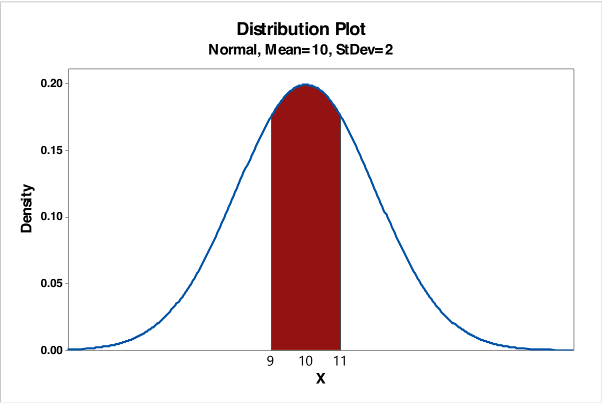


$$P(9 < X < 11) = P(X < 11) - P(X < 9)$$

$$= P\left(\frac{X-\mu}{\sigma} < \frac{11-\mu}{\sigma}\right) - P\left(\frac{X-\mu}{\sigma} < \frac{9-\mu}{\sigma}\right)$$

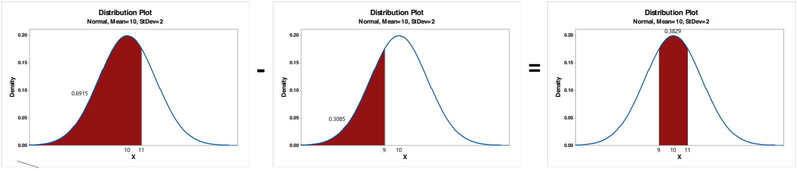
$$= P\left(Z < \frac{11-10}{2}\right) - P\left(Z < \frac{9-10}{2}\right) = P(Z < 0.5) - P(Z < -0.5)$$

$$= 0.6915 - 0.3085$$



Example: Normal Distribution

Suppose that the current measurements in a strip of wire are assumed to follow a normal distribution with μ = 10 and σ = 2 mA, what is the probability that the current measurement is between 9 and 11 mA?



Example: Normal Distribution

Suppose that the current measurements in a strip of wire are assumed to follow a normal distribution with μ = 10 and σ = 2 mA, what is the probability that the current measurement is between 9 and 11 mA?

```
pnorm(q=11, mean=10, sd=2) - pnorm(q=9, mean=10, sd=2)=
pnorm(q=0.5, mean=0, sd=1) - pnorm(q=-0.5, mean=0, sd=1)=
pnorm(q=0.5) - pnorm(q=-0.5)
```

Probability: 0.3829

Example: Normal Distribution determine percentiles …

Suppose that the current measurements in a strip of wire are assumed to follow a normal distribution with μ = 10 and σ = 2 mA.

Determine the value for which the probability that a current measurement is below 0.98.

Plot:

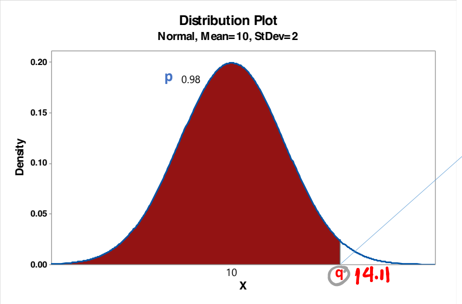
Example: Normal Distribution determine percentiles …

Determine the value for which the probability that a current measurement is below 0.98.

Plot:

$P(X<q)=0.98$

$q=?$



Supply p to find q

32

32

Example: Normal Distribution determine percentiles …

$$P(X<k)=0.98$$
$$P\left(\frac{X-10}{2}<\frac{k-10}{2}\right)=0.98$$
$$P\left(Z<\frac{k-10}{2}\right)=0.98$$

We also know from the normal table that:

~~$P(z<2.05)$~~ $= 0.98$

$P(Z<2.055)=0.98$

Therefore:

$P\left(Z<\frac{k-10}{2}\right)=P(Z<2.05)$ which means $\frac{k-10}{2}=2.055$

Then: $k=2*2.055+10=14.11$

$\boxed{2.05 \quad 2.06}$

0.9798 / 0.9803

0.98

$\frac{2.05+2.06}{2}=2.055$

$P(Z<2.055)=0.98$

33

33

Using R to calculate percentiles from a Normal Distribution

qnorm

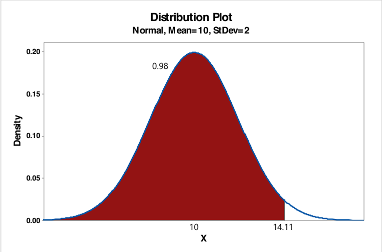
qnorm(p=0.??, mean= ??, sd=?? , lower.tail = ??)

qnorm is the inverse of the cdf, which you can also think of as the inverse of pnorm. Use qnorm to determine the x corresponding to the pth quantile of the normal distribution?

34

34

Determine the value for which the probability that a current measurement is below 0.98.



```
> qnorm(p=0.98, mean=10, sd=2, lower.tail = TRUE)
[1] 14.1075
```

≈ 14.11

35

35

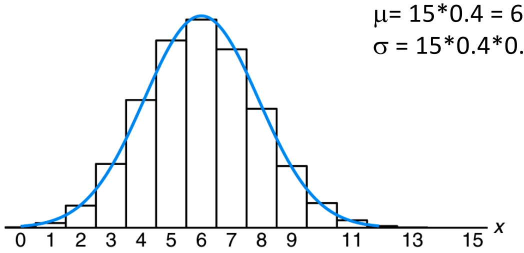
Normal Approximations

- The binomial and Poisson distributions become more bell-shaped and symmetric as their mean value increase.
  - If X ~Binomial(n, p) then X ~ $N(np, np(1-p))$
  - If X ~Poisson(λ) then X ~ $(N(\lambda, \lambda)$
- The normal distribution is a good approximation for:
  - Binomial if $np > 5$ and $n(1\text{-}p) > 5$. $X \sim B(n,p) \Rightarrow E(X) = np$ $Var(X) = np(1-p)$
  - Poisson if $\lambda > 5$. $X \sim P(\lambda) \Rightarrow E(X) = \lambda$ $Var(X) = \lambda$
- For manual calculations, the normal approximation is practical – use R for exact probabilities of the binomial and Poisson.

36

36

Normal approximation of *b*(*x*; n=15, p=0.4)



μ= 15*0.4 = 6,
σ = 15*0.4*0.6 = 3.6

37

Normal Approximation to the Poisson

If X is a Poisson random variable with $E(X) = \lambda$ and $V(X) = \lambda$,

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

The approximation is good for $\lambda \geq 5$

38

38

Continuity Correction

Using the normal distribution to approximate a discrete distribution (e.g. binomial) we need to take into account the fact that the normal distribution is continuous.

Discrete		Continuous
$P(X > k)$	$\longrightarrow$	$P\left(X > k + \frac{1}{2}\right)$
$P(X \geq k)$	$\longrightarrow$	$P\left(X > k - \frac{1}{2}\right)$
$P(X < k)$	$\longrightarrow$	$P\left(X < k - \frac{1}{2}\right)$
$P(X \leq k)$	$\longrightarrow$	$P\left(X < k + \frac{1}{2}\right)$
$P(k_1 < X < k_2)$	$\longrightarrow$	$P\left(k_1 + \frac{1}{2} < X < k_2 - \frac{1}{2}\right)$
$P(k_1 \leq X \leq k_2)$	$\longrightarrow$	$P\left(k_1 - \frac{1}{2} < X < k_2 + \frac{1}{2}\right)$

39

39

The number of phone calls at a call centre is Poisson distributed with mean 64 per hour. $X \sim P(\lambda = 64) \Rightarrow X \approx N(64, 64)$

1. What is the probability of 70 or more calls in a given hour?

2. What is the probability of less than 240 calls in a 4 hour period?

The number of phone calls at a call centre is Poisson distributed with mean 64 per hour.

1. What is the probability of 70 or more calls in a given hour? By using normal approximation to the poisson:

$$X \approx N(64, 64)$$

Pnorm(q = 69.5, mean = 64, sd = 64, Lower.tail = F)

$P(X \geq 70) = P(X > 70 - \frac{1}{2}) = P(X > 69.5) = P(\frac{X-64}{\sqrt{64}} > \frac{69.5-64}{\sqrt{64}}) = P(Z > 0.69) = 1 - P(Z < 0.69) = 1 - 0.7549 = 0.2451$

2. What is the probability of less than 240 calls in a 4 hour period? In four hours period $X_{4hrs} \sim \text{Poisson}(4 \times 64) = P(256)$

$$X_{4hrs} \approx N(4 \times 64, 4 \times 64) \equiv N(256, 256)$$

$P(X_{4hrs} < 240) = P(X_{4hrs} < 240 - \frac{1}{2}) = P(X_{4hrs} < 239.5) = P(\frac{X_{4hrs}-256}{\sqrt{256}} > \frac{239.5-256}{\sqrt{256}}) = P(Z < -1.03) = 0.1515$

8. Sampling distributions and confidence intervals

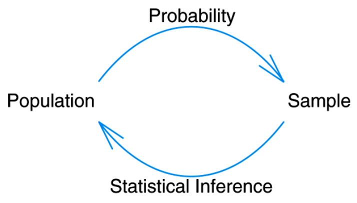
1

Learning Outcomes

- Explain sampling variation, sampling distribution, standard error
- Calculate the standard error of the sample mean
- State the Central Limit Theorem (applied to sampling distribution of the sample mean)
- Describe the sampling distribution of the sample mean in applications using the CLT
- Identify the point estimator of the parameter in applications
- Describe briefly the use of a confidence interval in inferential statistics
- Calculate and interpret 95% confidence interval for the population mean
- Use R to calculate the standard error and calculate a 95% confidence interval for the population mean
- Use the t distribution to calculate the standard error and confidence intervals for the population mean using a small sample
- Confidence intervals for the mean and other statistics via simulation, using R

2

Fundamental relationship between probability and inferential statistics



Probability

Population

Sample

Statistical Inference

3

Probability and Statistics

- In probability theory we consider some **known process** which has some randomness or uncertainty. We model the outcomes by random variables, and we figure out the probabilities of what will happen. There is one correct answer to any probability question.
- In statistical inference we observe something that has happened, and try to **figure out what underlying process** would explain those observations.

4

An example …

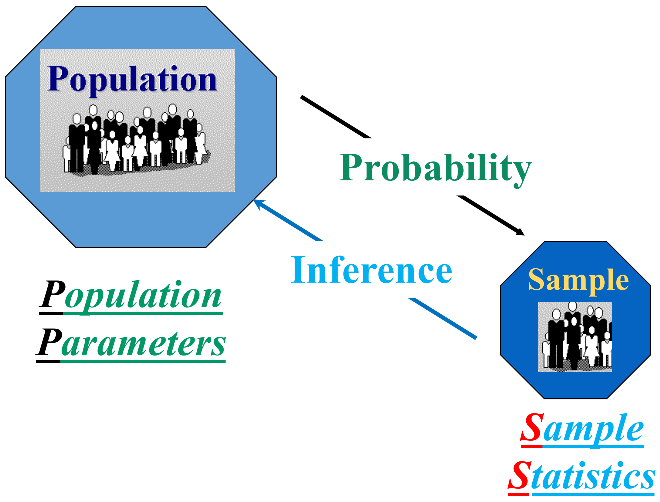
- Consider an (opaque) jar of red and green jelly beans.
- A probabilist starts by **knowing the proportion** of each and asks: What is the probability of drawing a red jelly bean from the jar?
- A statistician **infers the proportion** of red jelly beans by sampling from the jar, and using the sample proportion to estimate the jar proportion.

5

Probability and Statistics

- The basic aim behind all statistical methods is to make inferences about a population by studying a relatively small sample chosen from it.
- Probability is the engine that drives all statistical modelling, data analysis and inference.

6



Population

Probability

Inference

Sample

Population Parameters

Sample Statistics

7

Foundations for Inference

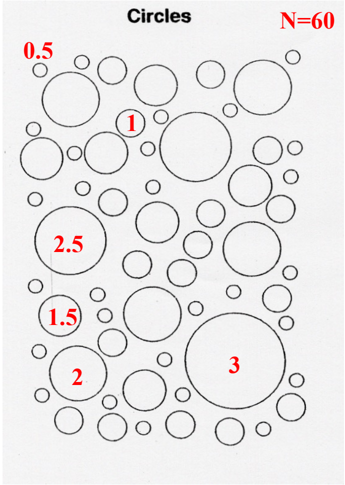
- Recall that inference is concerned (primarily) with estimating population parameters using sample statistics.
- A classic inferential question is, “How sure are we that the sample mean, x, is near the true population mean, µ?”
- Estimates (i.e. statistics) generally vary from one sample to another, and an understanding of **sampling variation** is key when estimating the precision of a sample statistic as an estimate of the corresponding parameter.

8

Sampling Distributions

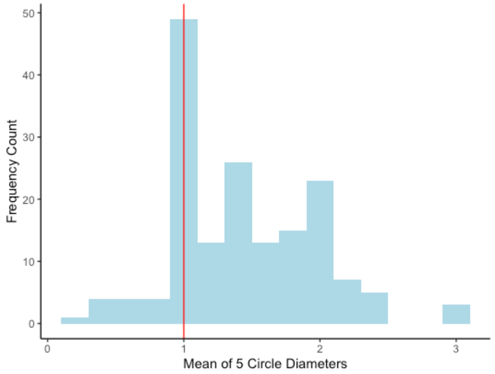
- The probability distribution of a statistic is called a sampling distribution.
- Sampling distributions arise because samples vary.
- Each random sample will have a different value of the statistic.

9

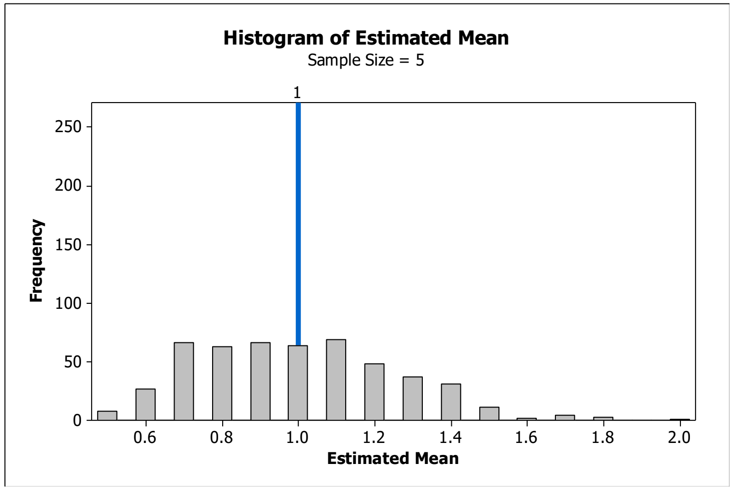


10

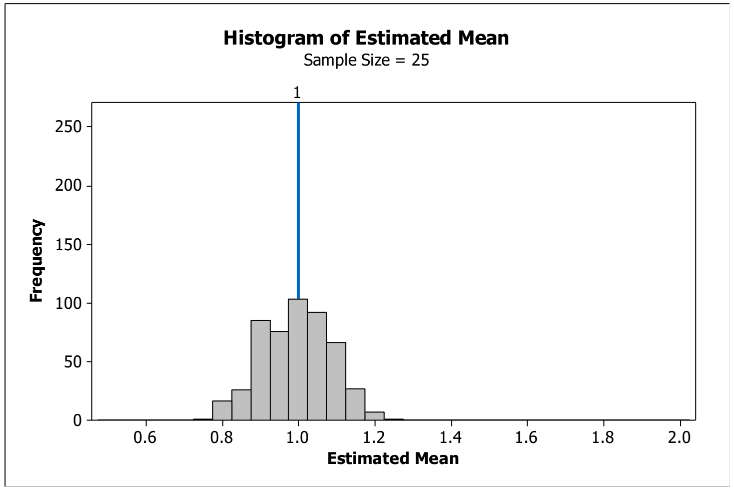
Judgement Sample



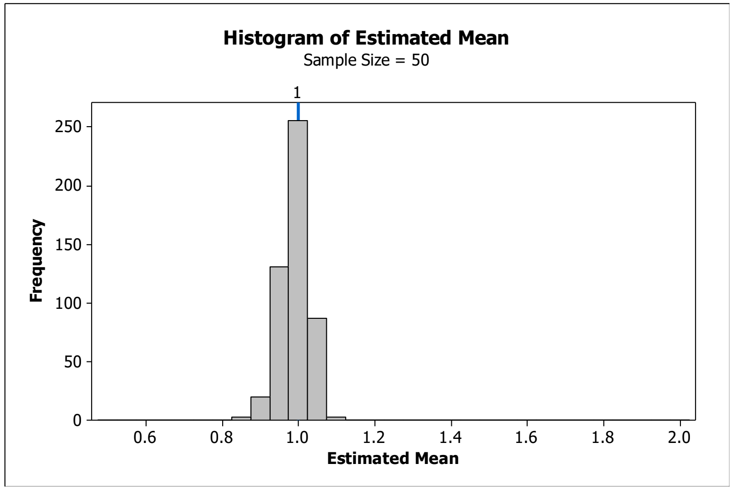
11



12

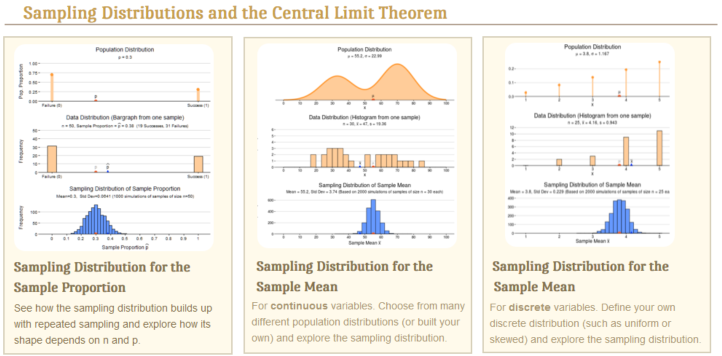


13



14

http://www.artofstat.com/webapps.html



Sampling Distributions and the Central Limit Theorem

Sampling Distribution for the Sample Proportion

See how the sampling distribution builds up with repeated sampling and explore how its shape depends on n and p.

Sampling Distribution for the Sample Mean

For **continuous** variables. Choose from many different population distributions (or built your own) and explore the sampling distribution.

Sampling Distribution for the Sample Mean

For **discrete** variables. Define your own discrete distribution (such as uniform or skewed) and explore the sampling distribution.

15

15

The Central Limit Theorem

- The sampling distribution of *any* mean becomes more nearly Normal as the sample size grows
  - observations need to be independent.
  - the shape of the population distribution doesn't matter.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

16

16

The Central Limit Theorem

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The CLT depends crucially on the assumption of independence.

You can't check this with your data. You have to think about how the data were gathered – can you assume the observations are independent?

The Central Limit Theorem

- *Sample means follow a Normal distribution centred on the population mean with a standard deviation equal to population standard deviation divided by the square root of the sample size.*

$$\boldsymbol{\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$$

- What happens when you take a single sample ?

The Standard Error

- The standard error is a measured of the variability in the sampling distribution (i.e. how do sample statistics vary about the unknown population parameter they are trying to estimate)
- It describes the typical 'error' or 'uncertainty' associated with the estimate.

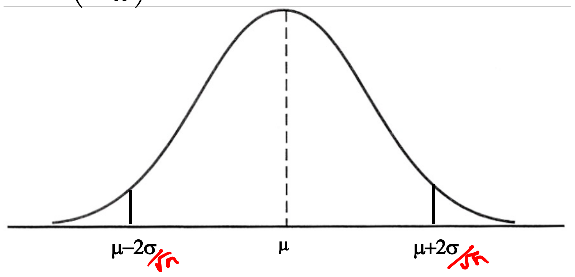
$$\boldsymbol{\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)} \qquad SE = \frac{\sigma}{\sqrt{n}}$$

Interval Estimation for μ

Use the CLT to provide a range of values that will capture 95% of sample means.

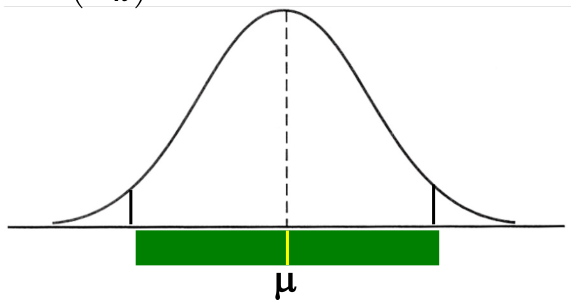
95% of sample means

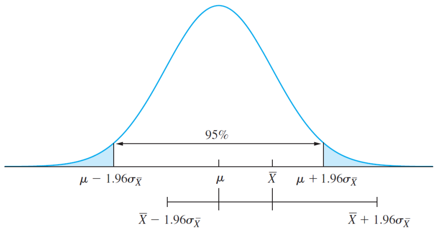
$$X \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



95% of sample means

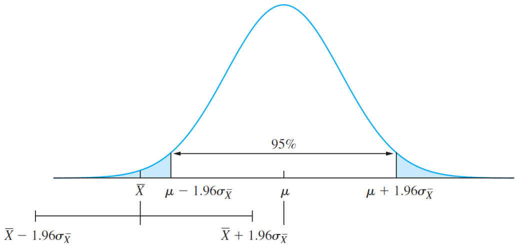
$$X \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$





95%
$\mu - 1.96\sigma_{\bar{X}}$
μ
$\bar{X}$
$\mu + 1.96\sigma_{\bar{X}}$
$\bar{X} - 1.96\sigma_{\bar{X}}$
$\bar{X} + 1.96\sigma_{\bar{X}}$

The sample mean X has a normal distribution with mean μ and standard deviation $\sigma_X = \sigma/\surd n$.
Let's consider a particular sample with mean x.
Now suppose x lies in the middle 95% of the distribution of X — the 95% confidence interval $x \pm 1.96\sigma_X$ succeeds in covering the population mean μ.



95%
$\bar{X}$
$\mu - 1.96\sigma_{\bar{X}}$
μ
$\mu + 1.96\sigma_{\bar{X}}$
$\bar{X} - 1.96\sigma_{\bar{X}}$
$\bar{X} + 1.96\sigma_{\bar{X}}$

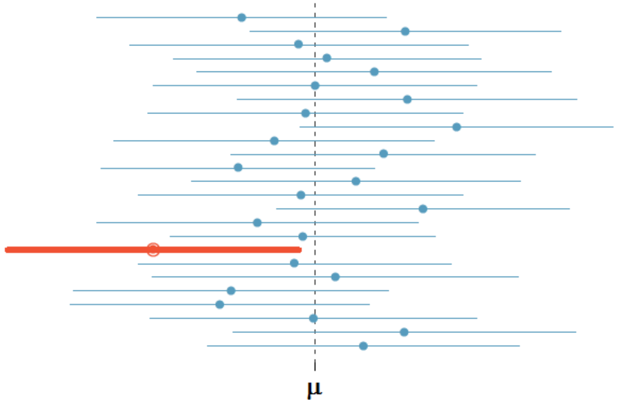
The sample mean X has a normal distribution with mean μ and standard deviation $\sigma_X = \sigma/\surd n$.
Let's consider a particular sample with mean x.
Now suppose x lies in the outer 5% of the distribution of X — the 95% confidence interval $x \pm 1.96\sigma_X$ does not include the population mean μ.

95% Confidence Interval for μ

In repeated sampling, 95% of intervals calculated in this manner

$$\bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

will contain the true mean μ.



Confidence intervals

- The population mean μ is **fixed**
- The intervals from different samples are **random**
- From our single sample, we only observe one of the intervals
- Our interval may or may not contain the true mean
- If we had taken many samples and calculated the 95% CI for each, 95% of them would include the true mean
- We say we are “95% confident” that the interval contains the true mean.

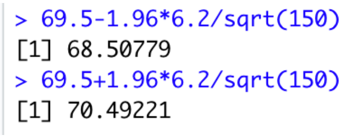
Confidence Intervals

- A point estimate (i.e. a statistic) is a single plausible value for a parameter.
- A point estimate is rarely perfect; usually there is some error in the estimate.
- Instead of supplying just a point estimate of a parameter, a next logical step would be to provide a plausible range of values for the parameter.
- To do this an estimate of the precision of the sample statistic (i.e. the estimate) is needed.

n = 150, $\bar{x}$ = 69.5, σ = 6.2

$$\boxed{\text{Is } \mu > 75\text{ ?}}$$

$$69.5 \pm 1.96 \frac{6.2}{\sqrt{150}}$$



```
> 69.5-1.96*6.2/sqrt(150)
[1] 68.50779
> 69.5+1.96*6.2/sqrt(150)
[1] 70.49221
```

A 95% CI for the population mean is [68.51, 70.49]

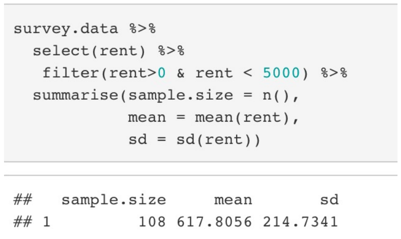
Interpret this !

Is μ > 75 ?

95% confident that the population mean is between 68.48 and 70.51 based on the data provided.

No evidence to support the claim that the population mean (μ) greater than 75.

Application: mean weekly rent in ST2001

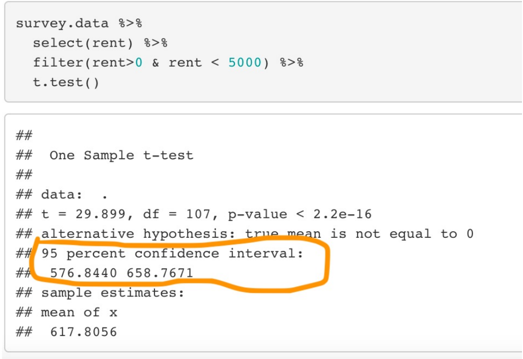


```
survey.data %>%
  select(rent) %>%
   filter(rent>0 & rent < 5000) %>%
  summarise(sample.size = n(),
            mean = mean(rent),
            sd = sd(rent))
```

```
##   sample.size     mean       sd
## 1         108 617.8056 214.7341
```

What is the population mean rent ?
What is a student likely to pay ?
What will they actually pay ?

Population Mean Rent in ST2001 ?



```
survey.data %>%
  select(rent) %>%
  filter(rent>0 & rent < 5000) %>%
  t.test()
```

```
##
##  One Sample t-test
##
## data:  .
## t = 29.899, df = 107, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  576.8440 658.7671
## sample estimates:
## mean of x
##  617.8056
```

Using s for σ ?

- Knowing s must mean that you knew μ ....

$$\bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

- The sample standard deviation s is used to estimate σ.
- What are the consequences ?

What if σ is unknown and n is small ?

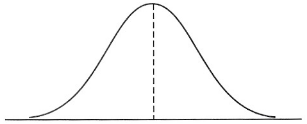
$$n < 30$$

$$\bar{x} \pm t_{(1-\frac{\alpha}{2}, n-1)} \frac{s}{\sqrt{n}}$$

1- confidence level

Degrees of free

Population normal



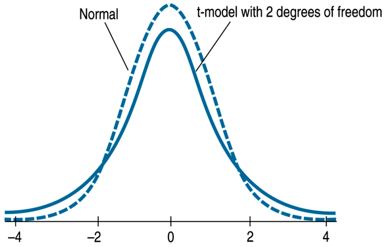
t_ν distribution

Mean = 0

Variance = $\frac{\nu}{\nu-2}$ **for** $\nu > 2$

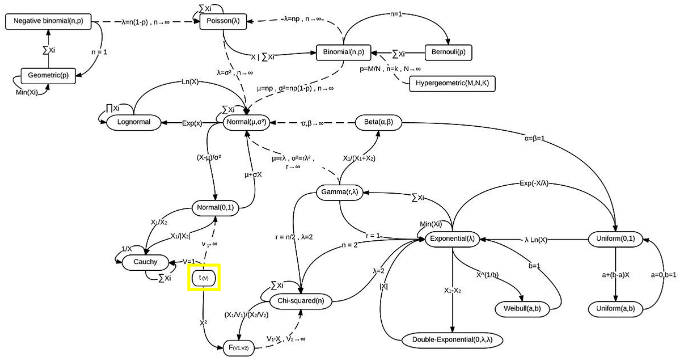
37

T- distribution

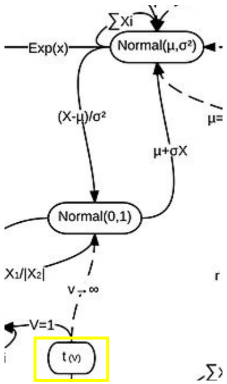


- As the degrees of freedom increase, the *t*-models look more and more like the Normal.
- In fact, the *t*-distribution with infinite degrees of freedom is the Normal distribution.

38

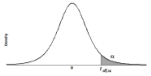


39



40

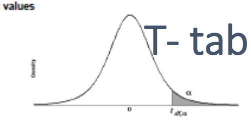
Table: t distribution critical values



T- tables

df	Upper tail probability											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
Z*	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291

Table: t distribution critical values



T- tables

df	Upper tail probability											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587

One-sample t-interval for a population mean

- When the conditions are met, we are ready to find the confidence interval for the population mean, μ.
- The confidence interval is

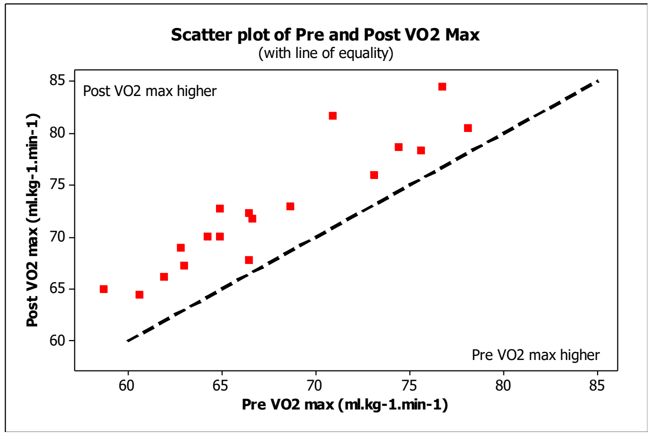
$$\bar{\boldsymbol{x}} \pm \boldsymbol{t}_{(1-\frac{\alpha}{2}, n-1)} \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}}$$

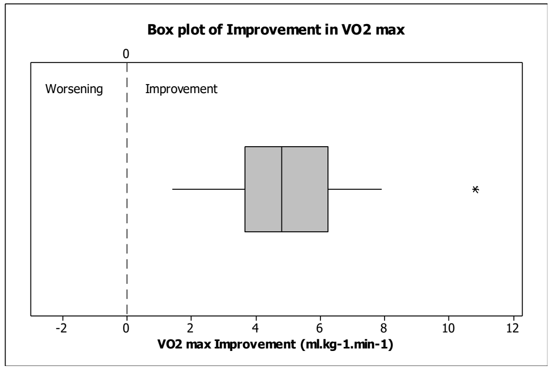
- The critical value $\boldsymbol{t}_{(1-\frac{\alpha}{2}, n-1)}$ depends on the particular confidence level, *1-* $\boldsymbol{\alpha}$, that you specify and on the number of degrees of freedom, *n* – 1, which we get from the sample size.
- **Let R do the work ….**

Example: Celtic study

Celtic Study

- A sample of 18 full-time youth soccer players from a Youth Academy performed high intensity aerobic interval training over a 10-week in-season period ***in addition*** to usual regime of soccer training and matches.
- Did this extra training improve fitness (VO2 max) ?
- Paired design: each player measured before and after (i.e. start and after 10 weeks)





```
Variable                      N     Mean     StDev
VO2 Improvement              18  5.11111   2.25829
```

Estimate the population mean improvement

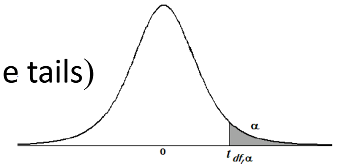
- 90% CI for μ
- 95% CI for μ
- 99% CI for μ

$$\bar{x} \pm t_{(1-\frac{\alpha}{2}, n-1)} \frac{s}{\sqrt{n}}$$

Estimate the population mean improvement

- 90% CI for μ
  ($\alpha = 0.10$ split over the tails)

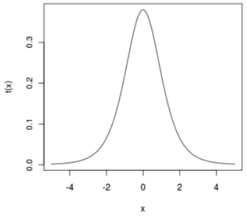
- 95% CI for μ
  ($\alpha = 0.05$ split over the tails)



- 99% CI for μ ($\alpha = 0.10$)
  $\alpha = 0.01$ split over the tails

Using R to calculate the quantile needed corresponding to a particular tail area

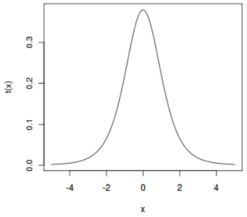
The qt(p=? , df= ?, lower.tail=TRUE) function calculates the t-value corresponding to a given lower-tailed area.



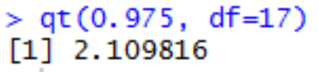
- Find the percentile of the Student t distribution needed for a 95% CI from a sample of size 18.

- Find the percentile of the Student t distribution needed for a 95% CI from a sample of size 18.

- For a 95% CI need the percentiles corresponding to tail areas such that 95% of the distribution is between these percentiles (i.e. 5% of the area split across the two tails).



- To calculate the 2.5th and 97.5th percentiles of the Student t distribution with 17 degrees of freedom:



```
> qt(0.975, df=17)
[1] 2.109816
```

Check the tables …

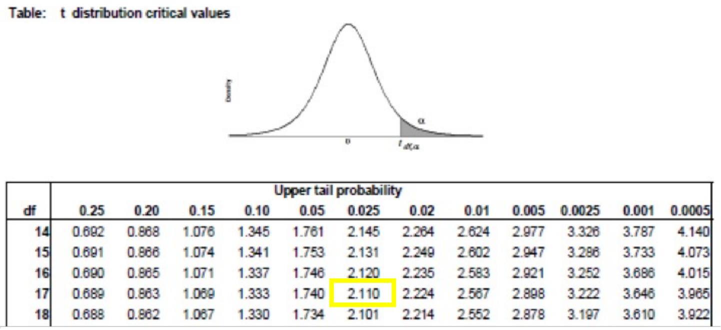


Table: t distribution critical values

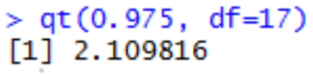
Density
α
0
$t_{df,\alpha}$

df	Upper tail probability											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922

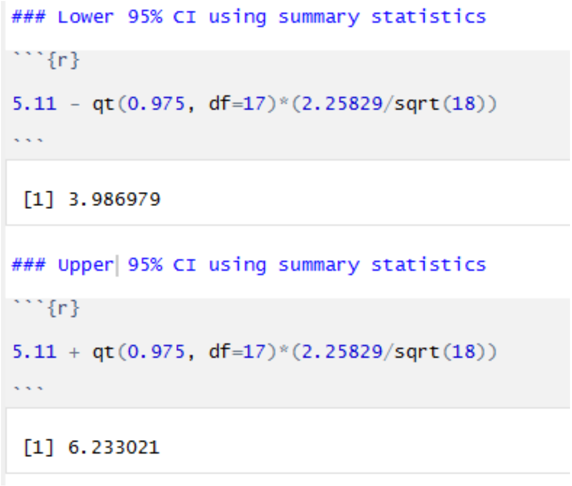
Estimate the population mean improvement

- 95% CI for μ $$\bar{x} \pm t_{(1-\frac{\alpha}{2}, n-1)} \frac{s}{\sqrt{n}}$$

```
Variable                   N     Mean    StDev
VO2 Improvement           18  5.11111  2.25829
```



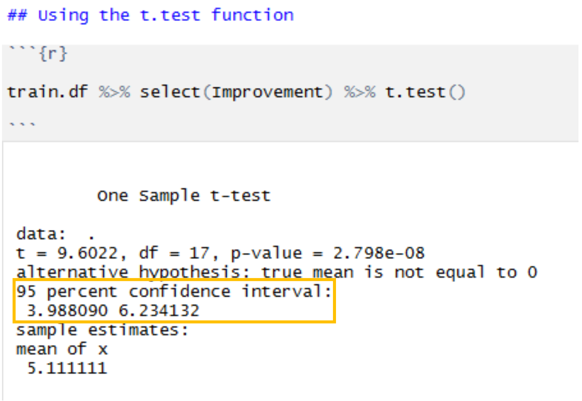
```
> qt(0.975, df=17)
[1] 2.109816
```



~~~
### Lower 95% CI using summary statistics
```{r}
5.11 - qt(0.975, df=17)*(2.25829/sqrt(18))
~~~

[1] 3.986979

Upper 95% CI using summary statistics
```{r}
5.11 + qt(0.975, df=17)*(2.25829/sqrt(18))
```

[1] 6.233021
```



```
## Using the t.test function
```{r}
train.df %>% select(Improvement) %>% t.test()
```

        One Sample t-test

data: .
t = 9.6022, df = 17, p-value = 2.798e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.988090 6.234132
sample estimates:
mean of x
 5.111111
```

Conclusion ?

- On average ?
- What does 95% Confidence mean ?
- Terms and conditions ?
- Random sample ?
- Small n, normality ??
```

# If Normality is questionable

a) Try to transform the data to approximate Normality

- e.g. logarithms or square root

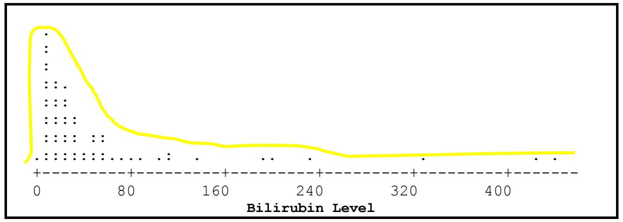
b) Non-Parametric technique

- Bootstrap
- CI for the population MEDIAN

57

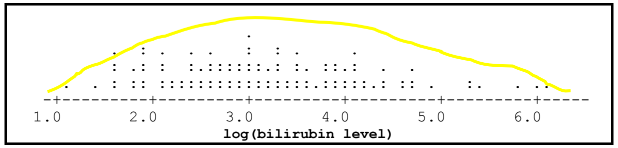
## Transforming to Normality

- ***Example:*** A study of Bilirubin levels in patients with Liver Disease



58

## Logarithm of Bilirubin Data



1. **Produce an interval estimate for the Population MEAN *log* bilirubin level**

2. **take anti-logs/exponentials of the resulting interval**

59

# If Normality is questionable

a) Try to transform the data to approximate Normality

- e.g. logarithms or square root

b) Non-Parametric technique

- Bootstrap
- CI for the population MEDIAN

60

## The Bootstrap

a) Try to transform the data to approximate Normality
   - e.g. logarithms or square root

b) Non-Parametric technique
   - Bootstrap
   - CI for the population MEDIAN

## Estimation via bootstrapping

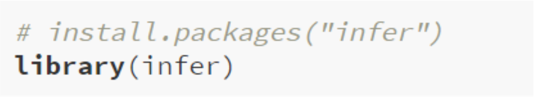
- We can quantify the variability of sample statistics using theory eg the Central Limit Theorem, or by simulation via bootstrapping.

- The term **bootstrapping** comes from the phrase "pulling oneself up by one’s bootstraps”.

## Bootstrapping scheme

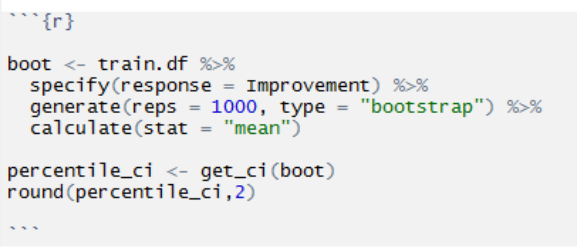
- Take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample.
- Calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples.
- Repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics.
- Calculate the bounds of the XX% confidence interval as the middle XX% of the bootstrap distribution.

## Bootstrapping in R



```
# install.packages("infer")
library(infer)
```

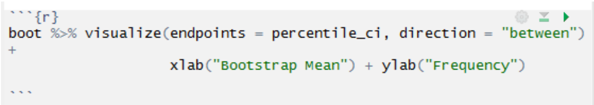
Generate bootstrap means



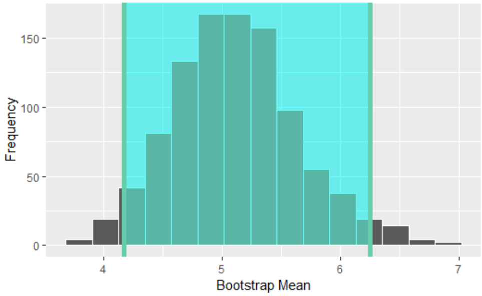
~~~
```{r}

boot <- train.df %>%
  specify(response = Improvement) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

percentile_ci <- get_ci(boot)
round(percentile_ci,2)

```
~~~

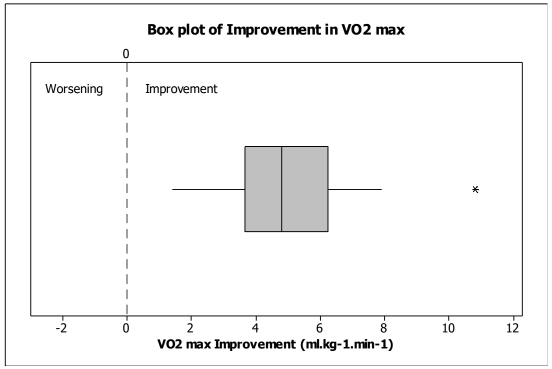
```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1   4.17    6.25
```

Plot the (empirical) sampling distribution



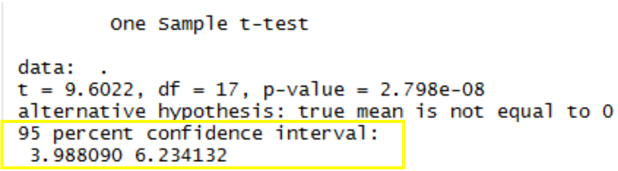
~~~
```{r}
boot %>% visualize(endpoints = percentile_ci, direction = "between")
+
                  xlab("Bootstrap Mean") + ylab("Frequency")
```
~~~



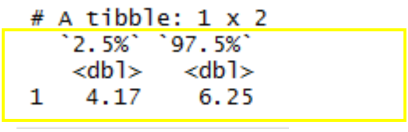


```
Variable                       N     Mean    StDev
VO2 Improvement               18  5.11111  2.25829
```

Compare the two 95% Confidence Intervals

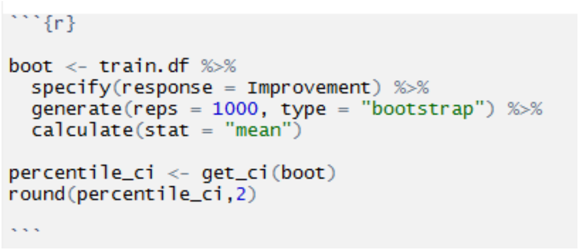


```
        One Sample t-test

data:  .
t = 9.6022, df = 17, p-value = 2.798e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.988090 6.234132
```



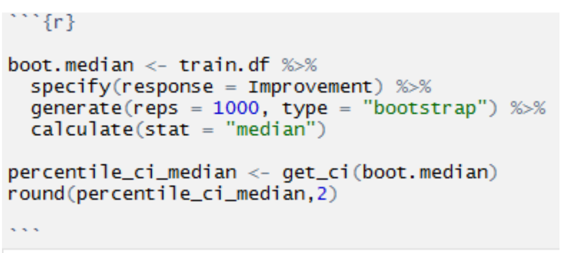
```
 # A tibble: 1 x 2
   `2.5%` `97.5%`
    <dbl>   <dbl>
 1   4.17    6.25
```

Generate bootstrap medians

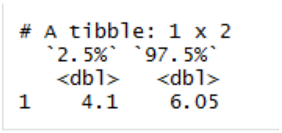


~~~
```{r}

boot <- train.df %>%
  specify(response = Improvement) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

percentile_ci <- get_ci(boot)
round(percentile_ci,2)

```
~~~

Generate bootstrap medians



~~~
```{r}

boot.median <- train.df %>%
  specify(response = Improvement) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "median")

percentile_ci_median <- get_ci(boot.median)
round(percentile_ci_median,2)

```
~~~



```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1    4.1    6.05
```

Celtic Study

- Based on the data provided the sample mean improvement was 5.11 mL/kg/min. We are 95% confident that the typical improvement in VO2 max is likely to be between 4 and 6 mL/kg/min.

- Given that the typical VO2 max at the start of this study was 67.66, the estimated typical improvement is approximately 7% (i.e. 5.11/67.66 expressed as percentage is 0.07*100).

- How would you translate this ?

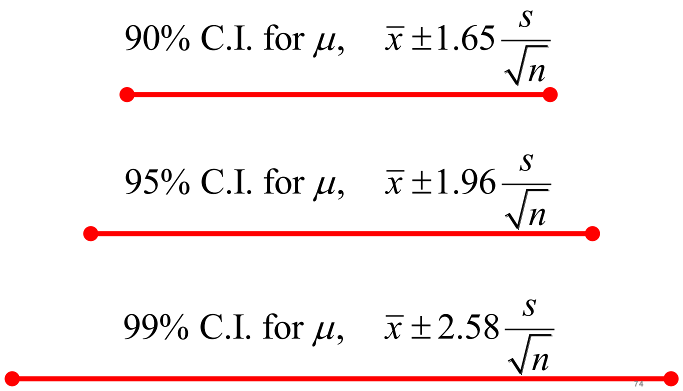
Celtic Study

- Does this mean that each player will improve by 5.11 units ?

Pick a parameter of interest ….

1. *Estimate it using an (unbiased) estimator*
2. *Calculate its corresponding standard error*;
3. *Calculate the corresponding (1-α)100% CI*;
4. Check the terms and conditions
5. Report the conclusions of the analysis.

Effect of increasing the confidence level

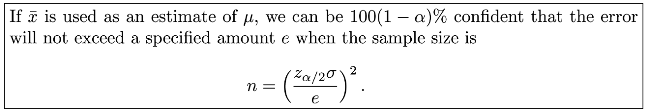


90% C.I. for μ, $\quad \bar{x} \pm 1.65 \frac{s}{\sqrt{n}}$

95% C.I. for μ, $\quad \bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$

99% C.I. for μ, $\quad \bar{x} \pm 2.58 \frac{s}{\sqrt{n}}$

Theorem 9.2



If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is

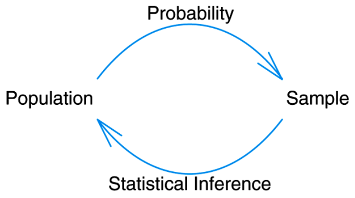
$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Very useful for sample size calculations

8. Sampling distributions and confidence intervals

1

Fundamental relationship between probability and inferential statistics



Probability

Population

Sample

Statistical Inference

2

2

Probability and Statistics

- In probability theory we consider some **known process** which has some randomness or uncertainty. We model the outcomes by random variables, and we figure out the probabilities of what will happen. There is one correct answer to any probability question.
- In statistical inference we observe something that has happened, and try to **figure out what underlying process** would explain those observations.

3

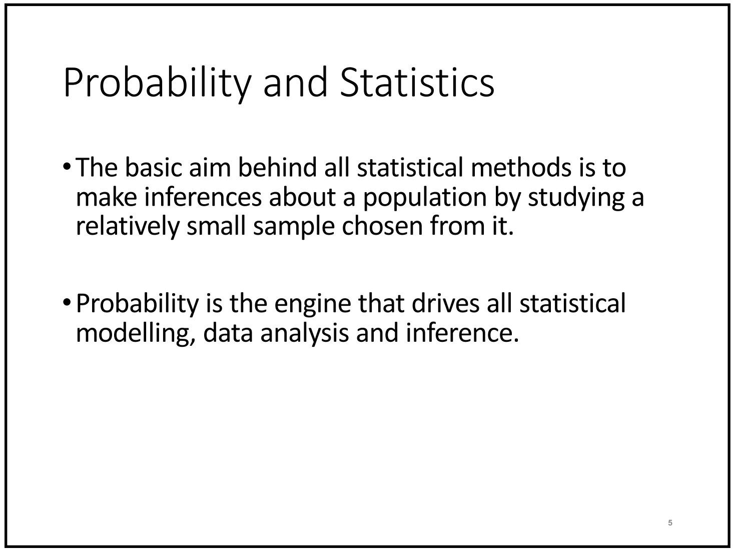
3

An example …

- Consider an (opaque) jar of red and green jelly beans.
- A probabilist starts by **knowing the proportion** of each and asks: What is the probability of drawing a red jelly bean from the jar?
- A statistician **infers the proportion** of red jelly beans by sampling from the jar, and using the sample proportion to estimate the jar proportion.

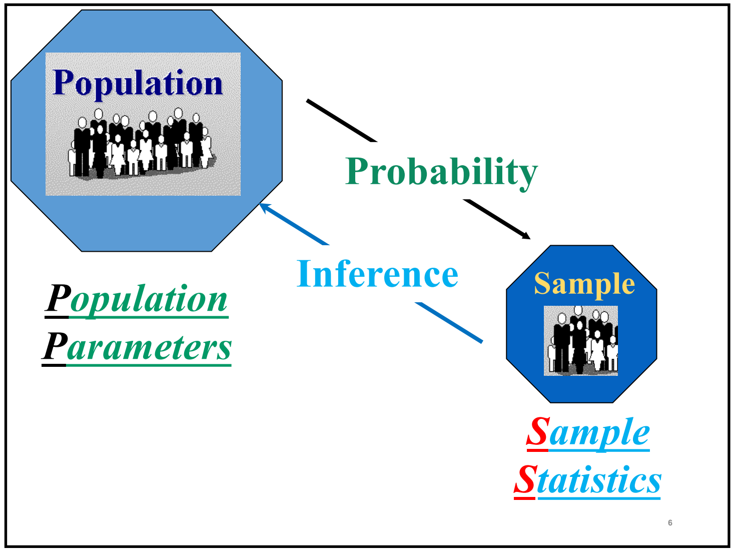
4

4



Probability and Statistics

- The basic aim behind all statistical methods is to make inferences about a population by studying a relatively small sample chosen from it.
- Probability is the engine that drives all statistical modelling, data analysis and inference.



Population

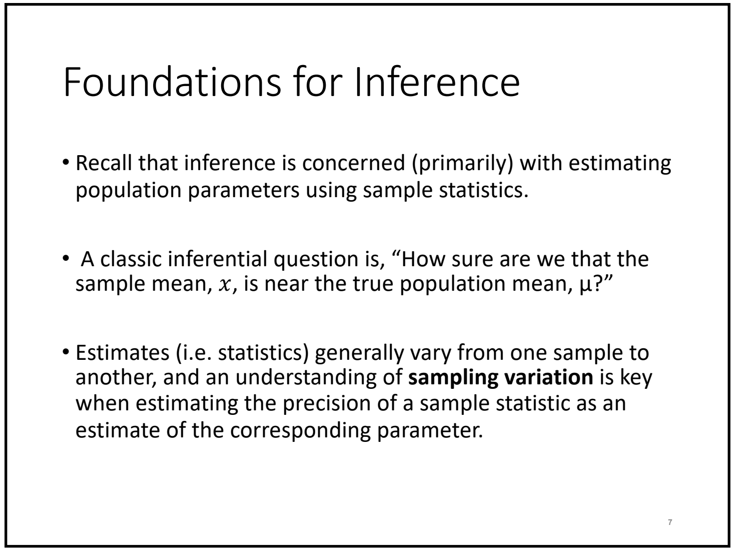
Probability

Inference

Sample

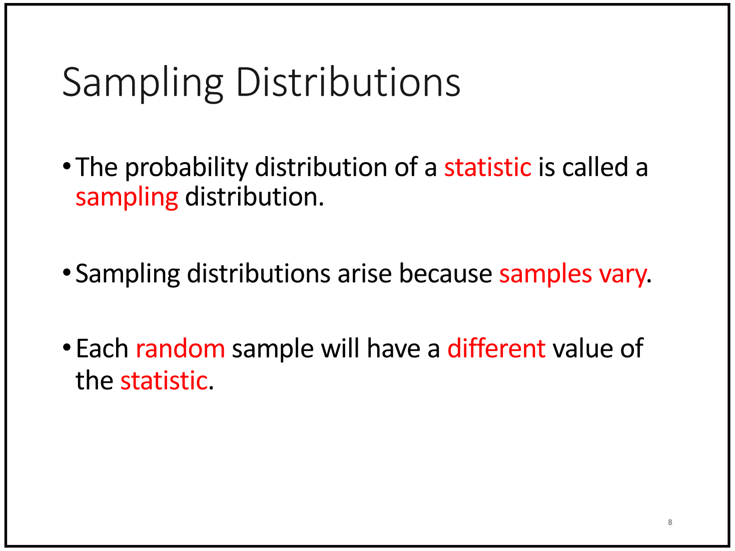
***P**opulation*
***P**arameters*

***S**ample*
***S**tatistics*



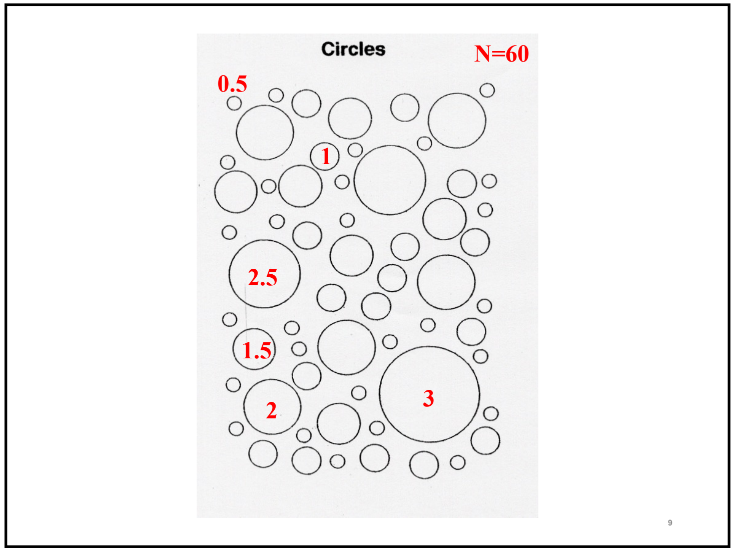
Foundations for Inference

- Recall that inference is concerned (primarily) with estimating population parameters using sample statistics.
- A classic inferential question is, “How sure are we that the sample mean, x, is near the true population mean, μ?”
- Estimates (i.e. statistics) generally vary from one sample to another, and an understanding of **sampling variation** is key when estimating the precision of a sample statistic as an estimate of the corresponding parameter.



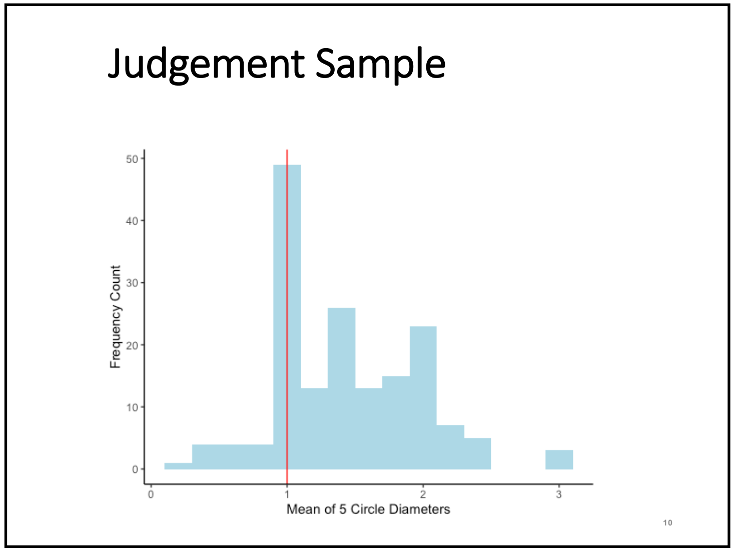
Sampling Distributions

- The probability distribution of a statistic is called a sampling distribution.
- Sampling distributions arise because samples vary.
- Each random sample will have a different value of the statistic.

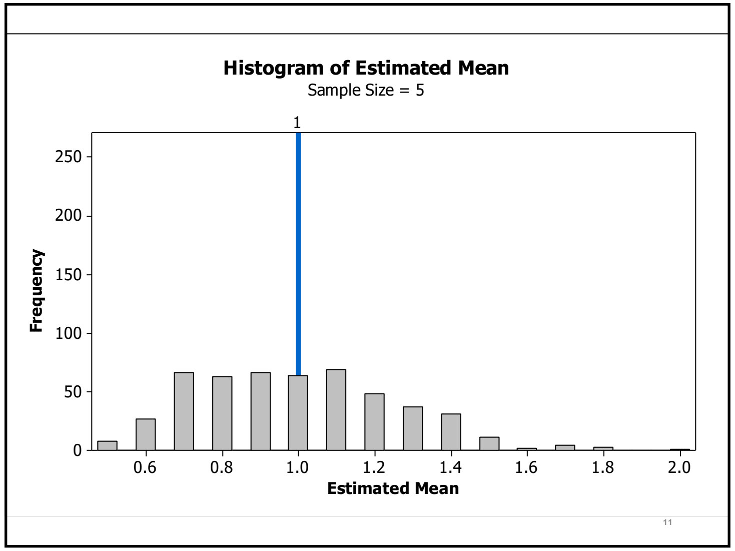


9

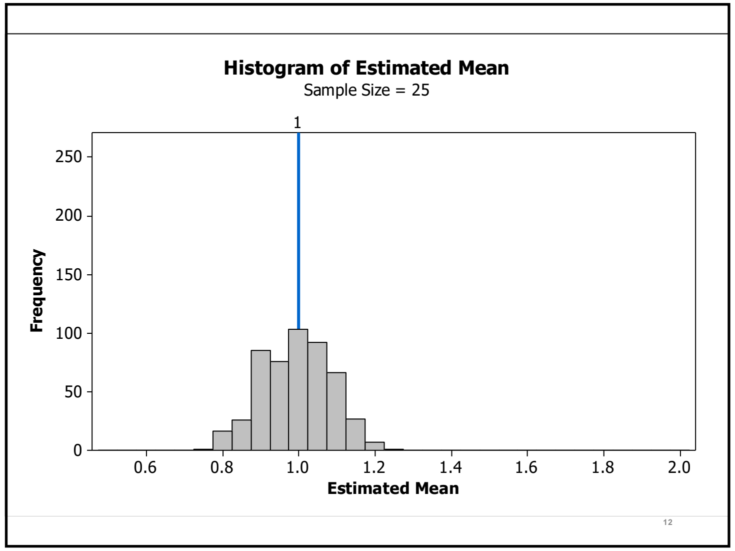
Judgement Sample



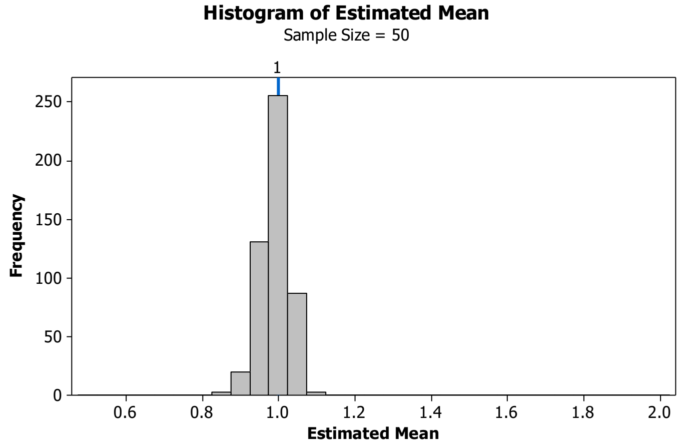
10



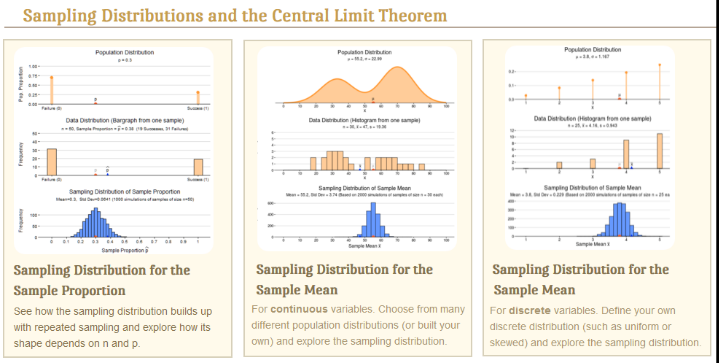
11



12



http://www.artofstat.com/webapps.html


Population Distribution
p = 0.3
Pop. Proportion
Failure (0)
Success (1)
Data Distribution (Bargraph from one sample)
n = 50, Sample Proportion = p̂ = 0.38 (19 Successes, 31 Failures)
Frequency
Sampling Distribution of Sample Proportion
Mean=0.3, Std Dev=0.0641 (1000 simulations of samples of size n=50)
Sample Proportion p̂
Population Distribution
μ = 55.2, σ = 22.99
Data Distribution (Histogram from one sample)
n = 30, x̄ = 47, s = 19.36
Sampling Distribution of Sample Mean
Mean = 55.2, Std Dev = 3.74 (Based on 2000 simulations of samples of size n = 30 each)
Sample Mean x̄
Population Distribution
μ = 3.8, σ = 1.167
Data Distribution (Histogram from one sample)
n = 25, x̄ = 4.16, s = 0.943
Sampling Distribution of Sample Mean
Mean = 3.8, Std Dev = 0.229 (Based on 2000 simulations of samples of size n = 25 ea
Sample Mean x̄

Sampling Distributions and the Central Limit Theorem

Sampling Distribution for the Sample Proportion

See how the sampling distribution builds up with repeated sampling and explore how its shape depends on n and p.

Sampling Distribution for the Sample Mean

For **continuous** variables. Choose from many different population distributions (or built your own) and explore the sampling distribution.

Sampling Distribution for the Sample Mean

For **discrete** variables. Define your own discrete distribution (such as uniform or skewed) and explore the sampling distribution.

The Central Limit Theorem

- The sampling distribution of *any* mean becomes more nearly Normal as the sample size grows
  - observations need to be independent.
  - the shape of the population distribution doesn't matter.

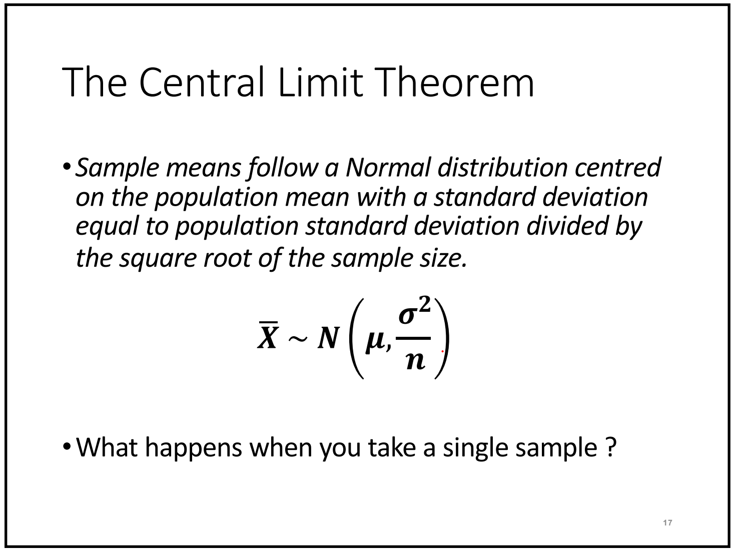
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The Central Limit Theorem

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The CLT depends crucially on the assumption of independence.

You can’t check this with your data. You have to think about how the data were gathered – can you assume the observations are independent?



The Central Limit Theorem

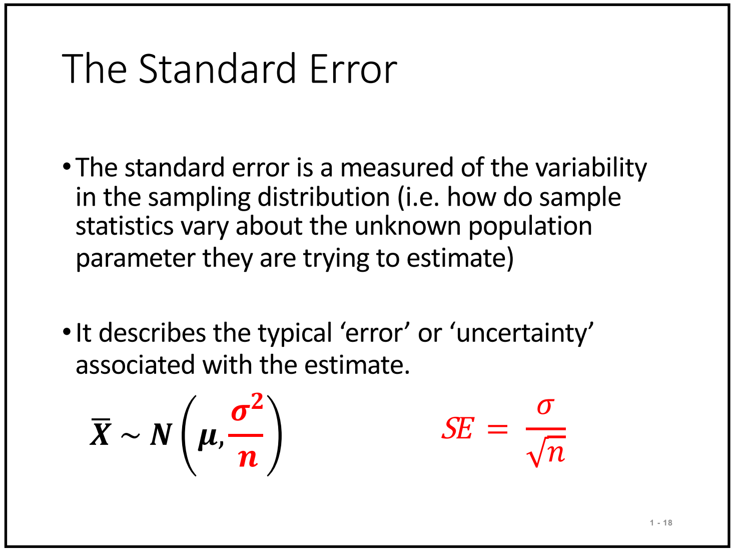
- *Sample means follow a Normal distribution centred on the population mean with a standard deviation equal to population standard deviation divided by the square root of the sample size.*

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- What happens when you take a single sample ?

17

17



The Standard Error

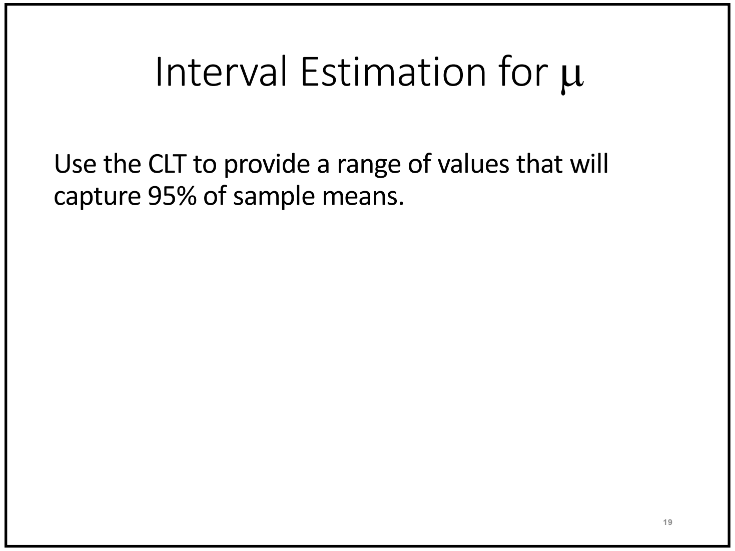
- The standard error is a measured of the variability in the sampling distribution (i.e. how do sample statistics vary about the unknown population parameter they are trying to estimate)

- It describes the typical 'error' or 'uncertainty' associated with the estimate.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad SE = \frac{\sigma}{\sqrt{n}}$$

1 - 18

18

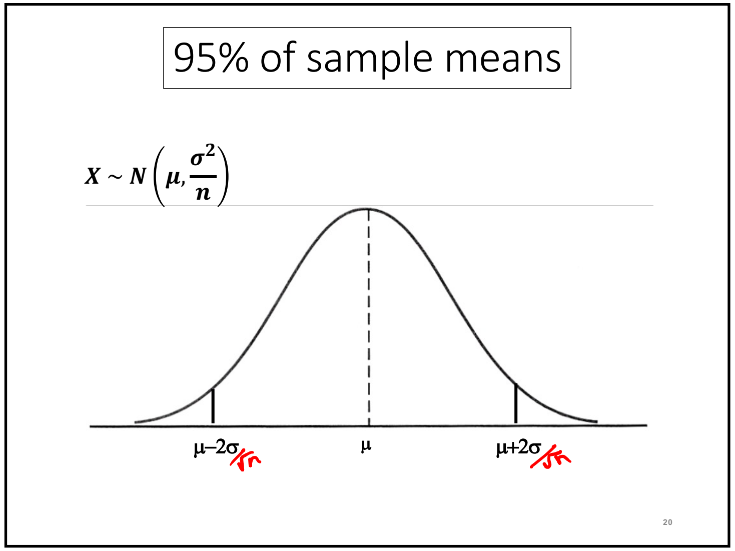


Interval Estimation for μ

Use the CLT to provide a range of values that will capture 95% of sample means.

19

19

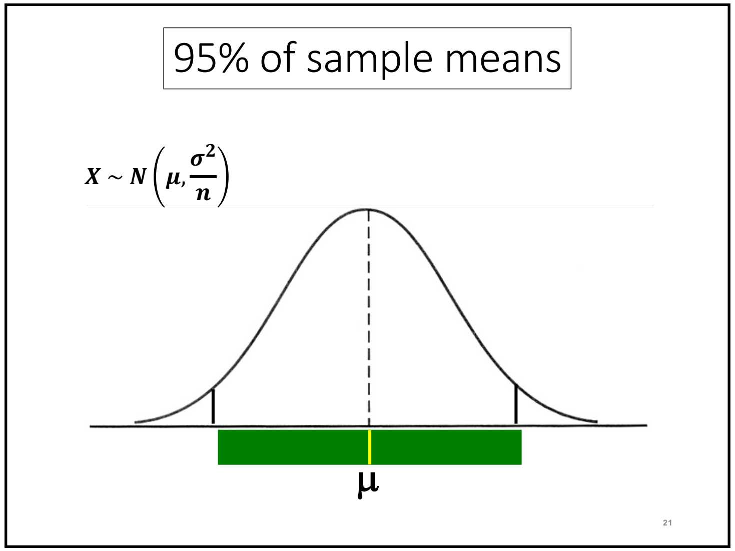


95% of sample means

$$X \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

20

20

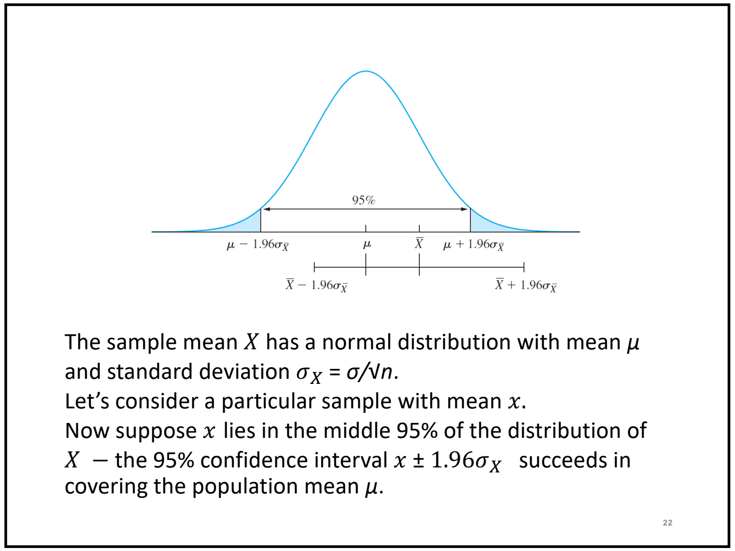


95% of sample means

$X \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

21

21

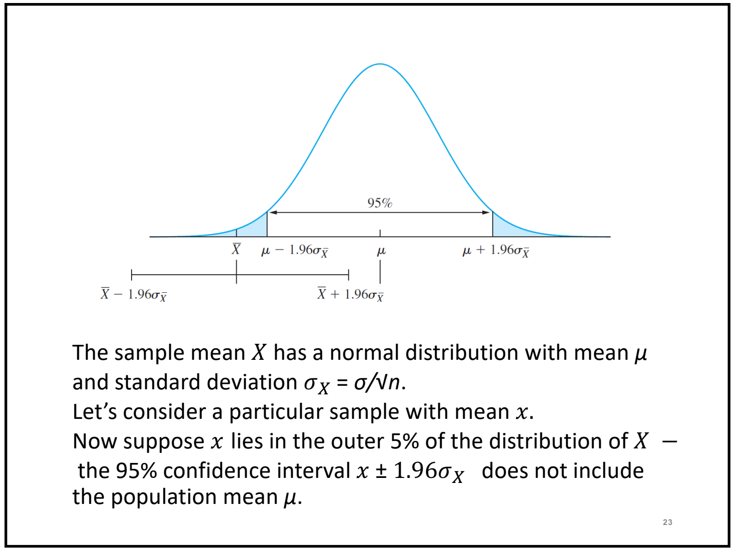


95%
$\mu - 1.96\sigma_{\bar{X}}$
μ
$\bar{X}$
$\mu + 1.96\sigma_{\bar{X}}$
$\bar{X} - 1.96\sigma_{\bar{X}}$
$\bar{X} + 1.96\sigma_{\bar{X}}$

The sample mean X has a normal distribution with mean μ and standard deviation $\sigma_X = \sigma/\sqrt{n}$.
Let's consider a particular sample with mean x.
Now suppose x lies in the middle 95% of the distribution of X – the 95% confidence interval $x \pm 1.96\sigma_X$ succeeds in covering the population mean μ.

22

22

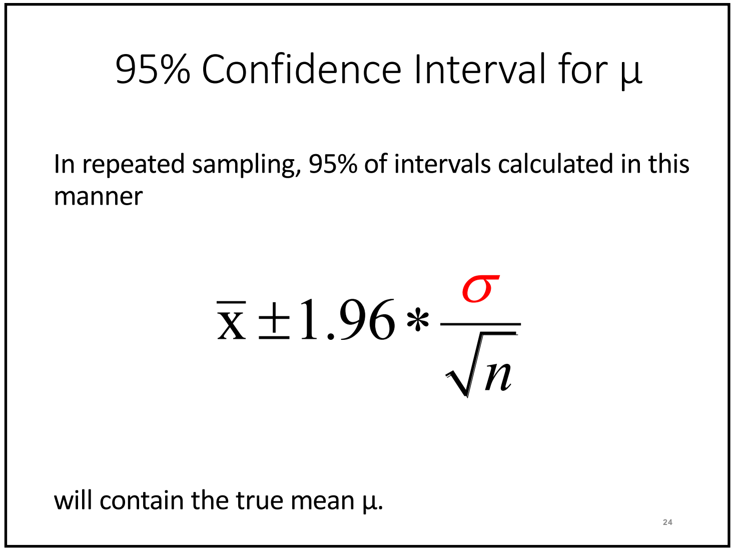


95%
$\bar{X}$
$\mu - 1.96\sigma_{\bar{X}}$
μ
$\mu + 1.96\sigma_{\bar{X}}$
$\bar{X} - 1.96\sigma_{\bar{X}}$
$\bar{X} + 1.96\sigma_{\bar{X}}$

The sample mean X has a normal distribution with mean μ and standard deviation $\sigma_X = \sigma/\sqrt{n}$.
Let's consider a particular sample with mean x.
Now suppose x lies in the outer 5% of the distribution of X – the 95% confidence interval $x \pm 1.96\sigma_X$ does not include the population mean μ.

23

23



95% Confidence Interval for μ

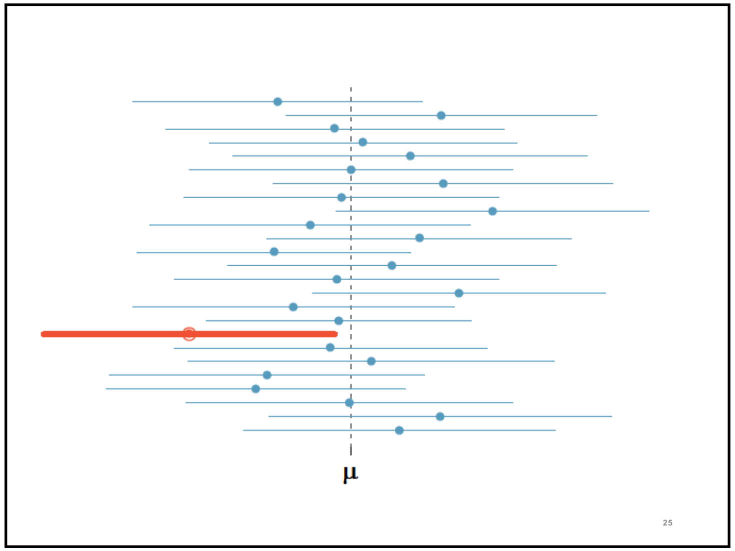
In repeated sampling, 95% of intervals calculated in this manner

$$\bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

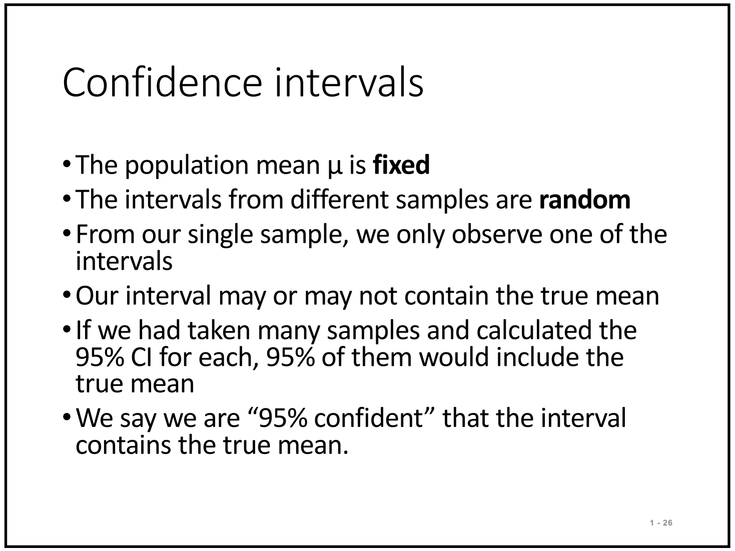
will contain the true mean μ.

24

24



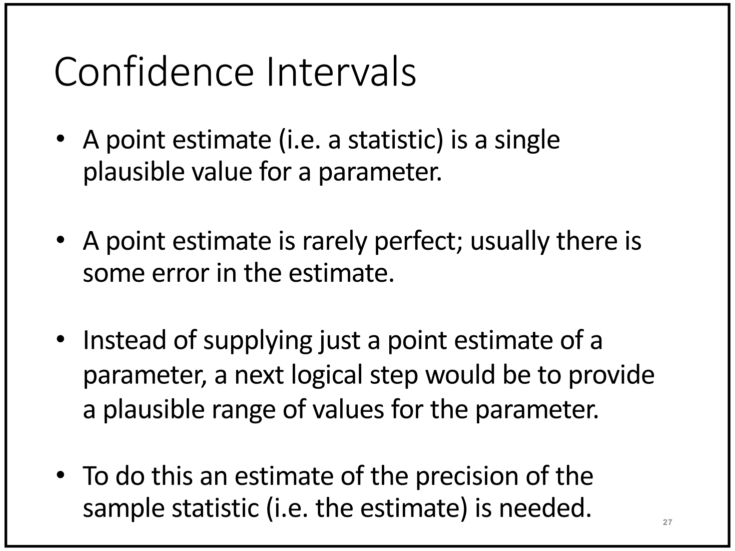
25



Confidence intervals

- The population mean μ is **fixed**
- The intervals from different samples are **random**
- From our single sample, we only observe one of the intervals
- Our interval may or may not contain the true mean
- If we had taken many samples and calculated the 95% CI for each, 95% of them would include the true mean
- We say we are “95% confident” that the interval contains the true mean.

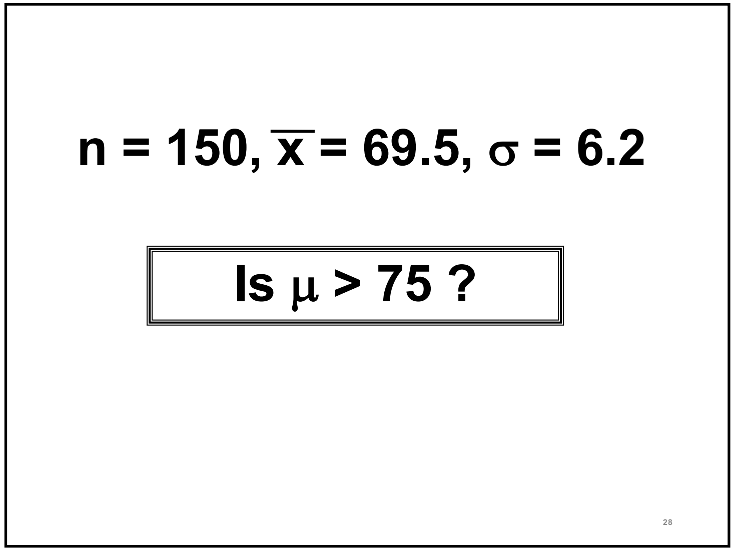
26



Confidence Intervals

- A point estimate (i.e. a statistic) is a single plausible value for a parameter.
- A point estimate is rarely perfect; usually there is some error in the estimate.
- Instead of supplying just a point estimate of a parameter, a next logical step would be to provide a plausible range of values for the parameter.
- To do this an estimate of the precision of the sample statistic (i.e. the estimate) is needed.

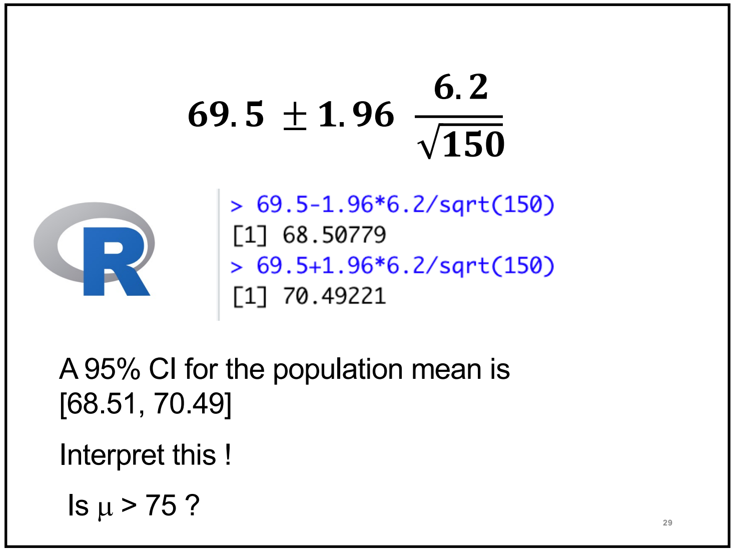
27



$$n = 150, \bar{x} = 69.5, \sigma = 6.2$$

$$\boxed{\text{Is } \mu > 75 \text{ ?}}$$

28



$$69.5 \pm 1.96 \frac{6.2}{\sqrt{150}}$$

```
> 69.5-1.96*6.2/sqrt(150)
[1] 68.50779
> 69.5+1.96*6.2/sqrt(150)
[1] 70.49221
```

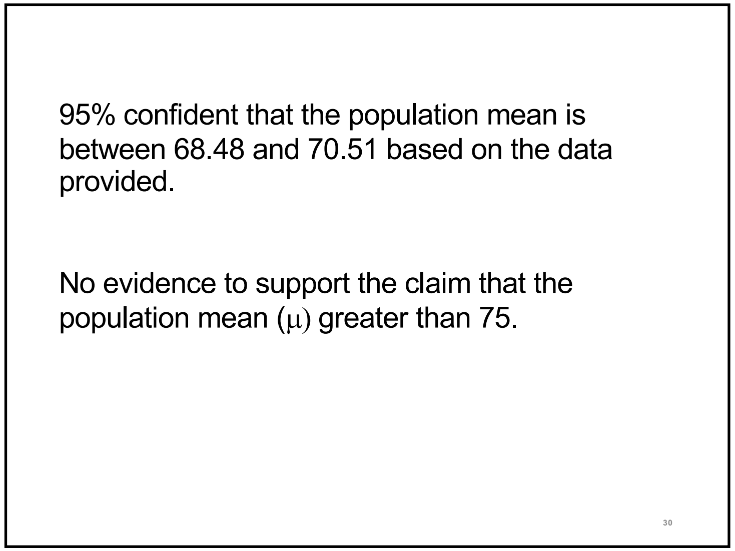
A 95% CI for the population mean is [68.51, 70.49]

Interpret this !

Is $\mu > 75$?

29

29

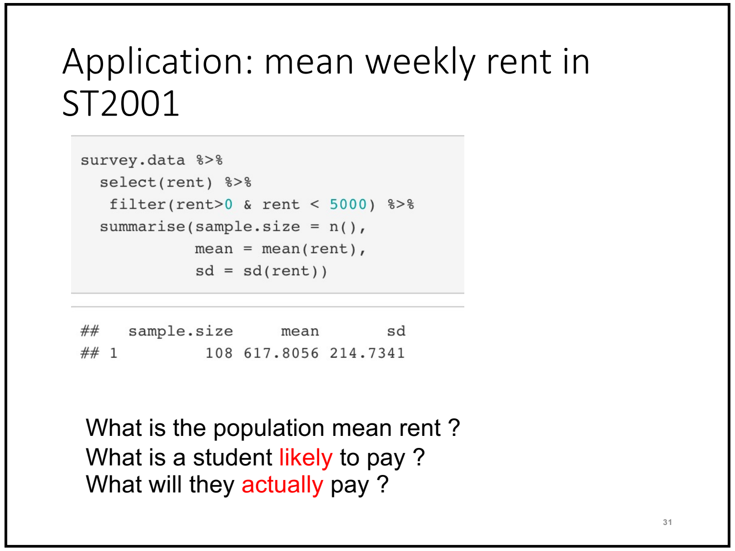


95% confident that the population mean is between 68.48 and 70.51 based on the data provided.

No evidence to support the claim that the population mean (μ) greater than 75.

30

30



Application: mean weekly rent in ST2001

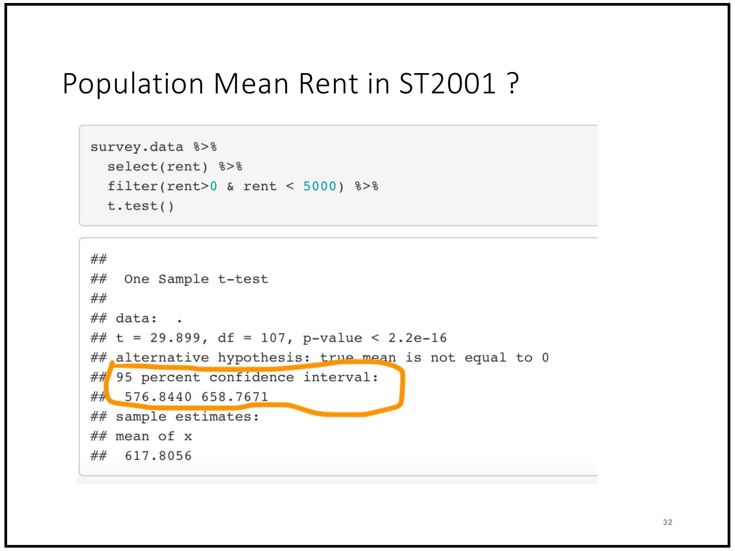
```
survey.data %>%
  select(rent) %>%
   filter(rent>0 & rent < 5000) %>%
  summarise(sample.size = n(),
            mean = mean(rent),
            sd = sd(rent))
```

```
##   sample.size     mean       sd
## 1         108 617.8056 214.7341
```

What is the population mean rent ?
What is a student likely to pay ?
What will they actually pay ?

31

31



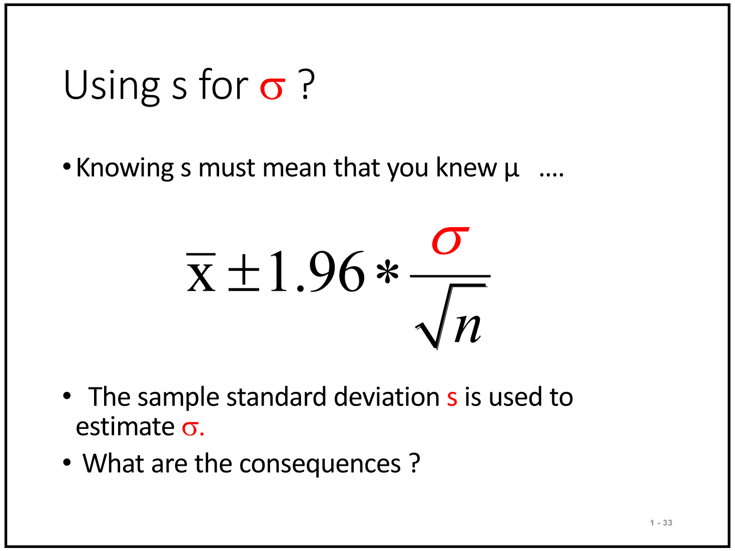
Population Mean Rent in ST2001 ?

```
survey.data %>%
  select(rent) %>%
  filter(rent>0 & rent < 5000) %>%
  t.test()
```

```
##
##  One Sample t-test
##
## data:  .
## t = 29.899, df = 107, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  576.8440 658.7671
## sample estimates:
## mean of x
##  617.8056
```

32

32



Using s for σ ?

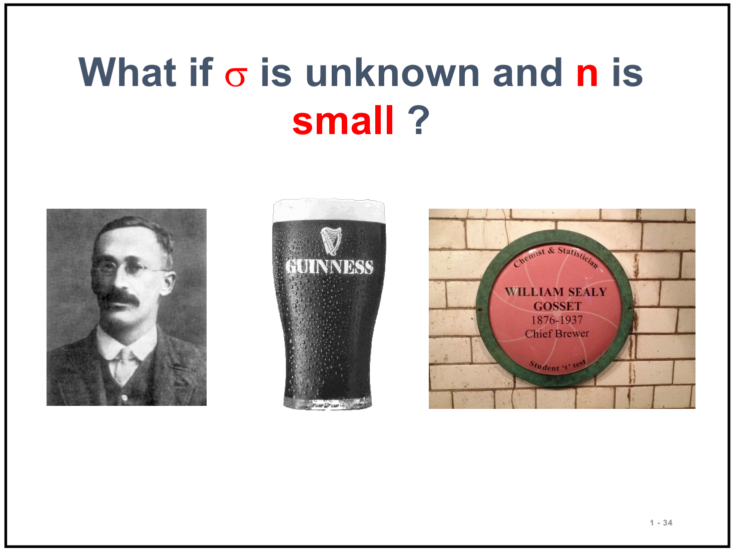
- Knowing s must mean that you knew μ ....

$$\bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

- The sample standard deviation s is used to estimate σ.
- What are the consequences ?

1 - 33

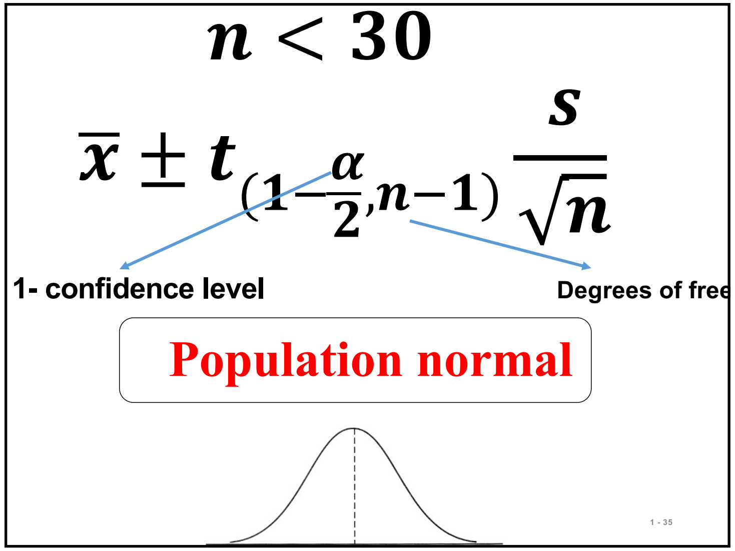
33



What if σ is unknown and n is small ?

1 - 34

34



$$n < 30$$

$$\bar{x} \pm t_{(1-\frac{\alpha}{2}, n-1)} \frac{s}{\sqrt{n}}$$

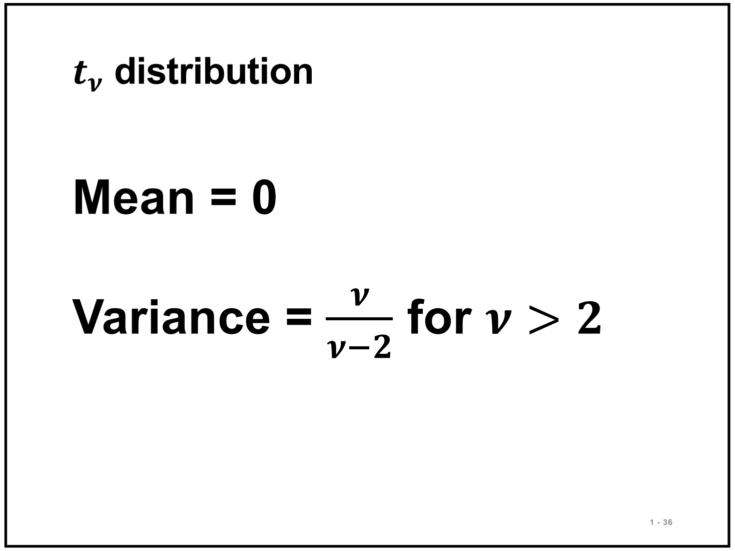
1- confidence level

Degrees of free

Population normal

1 - 35

35



t_ν distribution

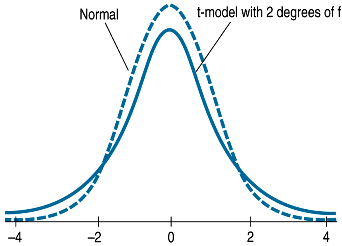
Mean = 0

Variance = $\frac{\nu}{\nu-2}$ for $\nu > 2$

1 - 36

36

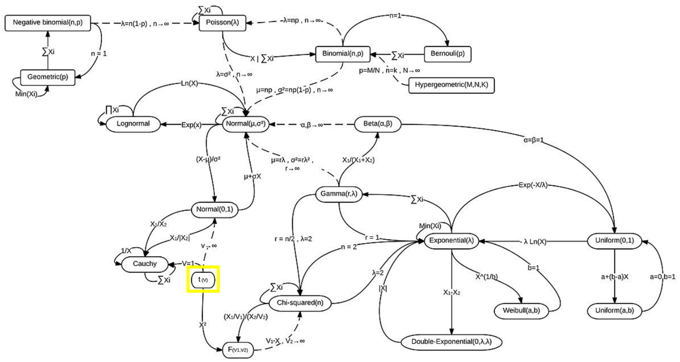
T- distribution



- As the degrees of freedom increase, the *t*-models look more and more like the Normal.
- In fact, the *t*-distribution with infinite degrees of freedom is the Normal distribution.

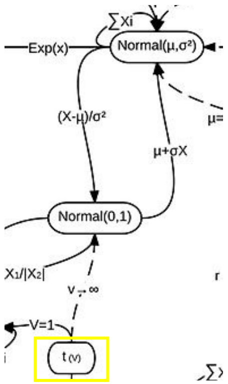
37

37



38

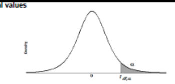
38



39

39

Table: t distribution critical values



T- tables

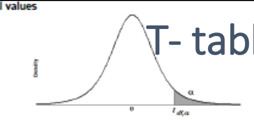
	Upper tail probability											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
Z*	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291

40

40

Table: t distribution critical values

T- tables



df	Upper tail probability											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587

One-sample t-interval for a population mean

- When the conditions are met, we are ready to find the confidence interval for the population mean, μ.
- The confidence interval is

$$\overline{\boldsymbol{x}} \pm \boldsymbol{t}_{\left(\mathbf{1}-\frac{\boldsymbol{\alpha}}{\mathbf{2}},\boldsymbol{n}-\mathbf{1}\right)} \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}}$$

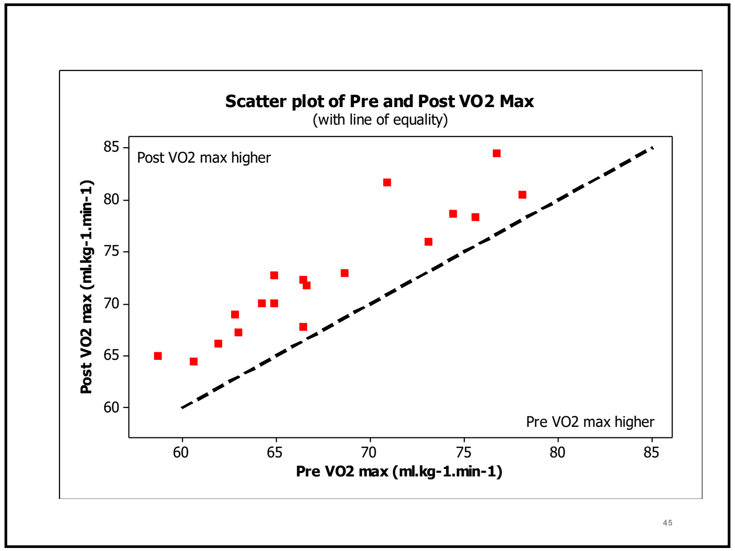
- The critical value $\boldsymbol{t}_{\left(1-\frac{\alpha}{2},n-1\right)}$ depends on the particular confidence level, *1*- $\boldsymbol{\alpha}$, that you specify and on the number of degrees of freedom, *n* – 1, which we get from the sample size.

- **Let R do the work ….**

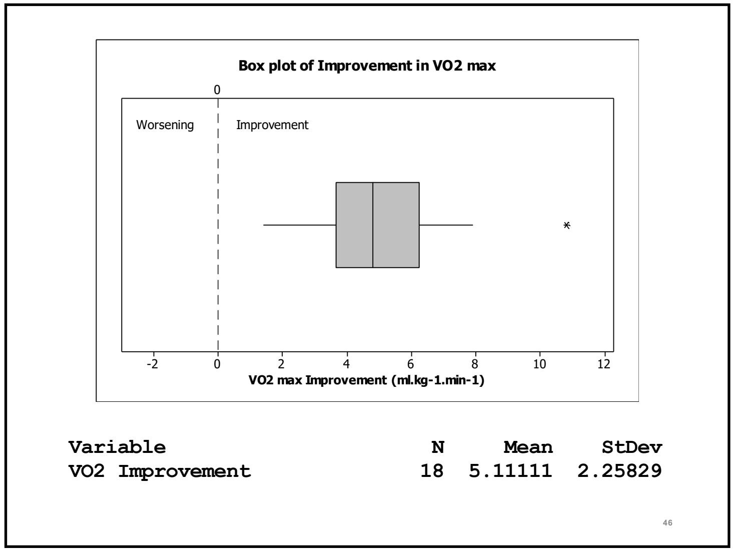
Example: Celtic study

Celtic Study

- A sample of 18 full-time youth soccer players from a Youth Academy performed high intensity aerobic interval training over a 10-week in-season period ***in addition*** to usual regime of soccer training and matches.
- Did this extra training improve fitness (VO2 max) ?
- Paired design: each player measured before and after (i.e. start and after 10 weeks)

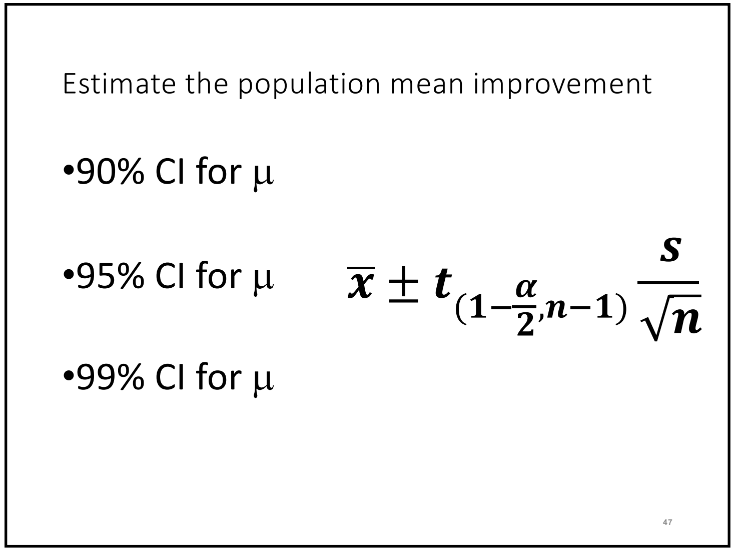


45



```
Variable              N    Mean    StDev
VO2 Improvement      18  5.11111  2.25829
```

46



Estimate the population mean improvement

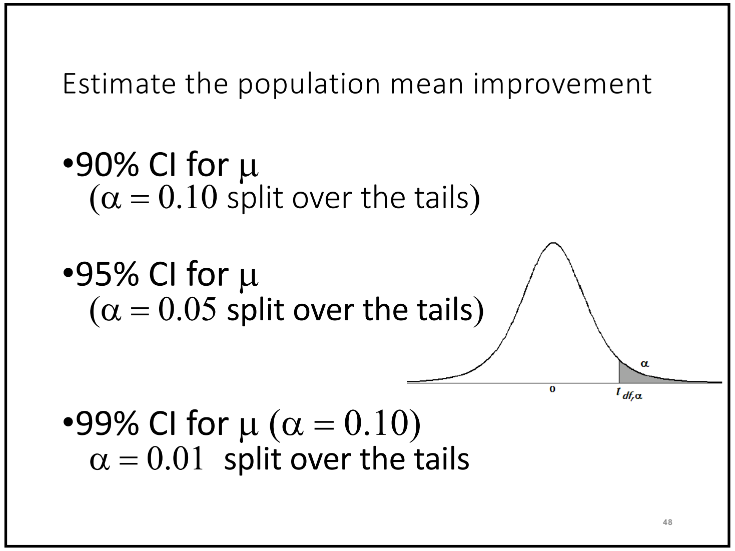
- 90% CI for μ

- 95% CI for μ

$$\bar{x} \pm t_{(1-\frac{\alpha}{2}, n-1)} \frac{s}{\sqrt{n}}$$

- 99% CI for μ

47



Estimate the population mean improvement

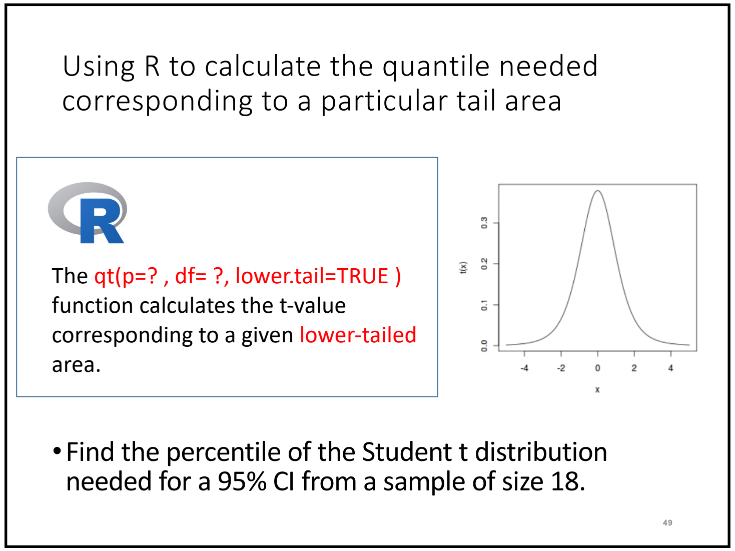
- 90% CI for μ
  (α = 0.10 split over the tails)

- 95% CI for μ
  (α = 0.05 split over the tails)

α
0
$t_{df,\alpha}$

- 99% CI for μ (α = 0.10)
  α = 0.01 split over the tails

48

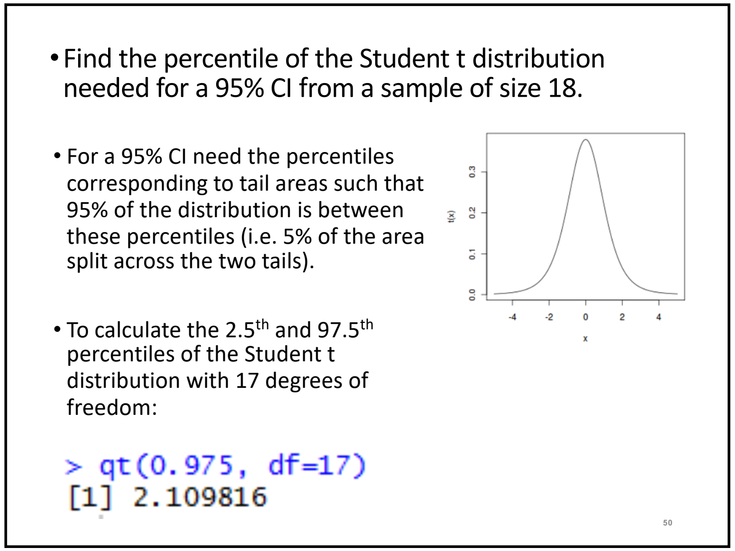


Using R to calculate the quantile needed corresponding to a particular tail area

The qt(p=? , df= ?, lower.tail=TRUE) function calculates the t-value corresponding to a given lower-tailed area.

- Find the percentile of the Student t distribution needed for a 95% CI from a sample of size 18.

49



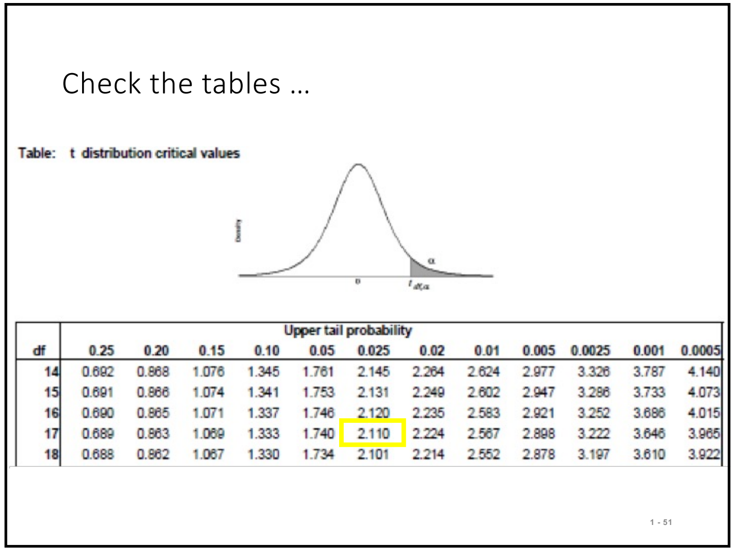
- Find the percentile of the Student t distribution needed for a 95% CI from a sample of size 18.

- For a 95% CI need the percentiles corresponding to tail areas such that 95% of the distribution is between these percentiles (i.e. 5% of the area split across the two tails).

- To calculate the 2.5th and 97.5th percentiles of the Student t distribution with 17 degrees of freedom:

```
> qt(0.975, df=17)
[1] 2.109816
```

50

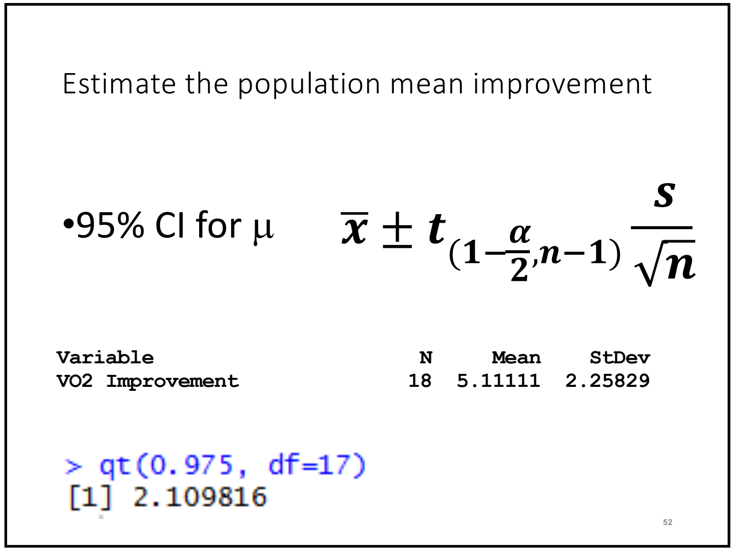


Check the tables …

Table: t distribution critical values

df	Upper tail probability											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922

51



Estimate the population mean improvement

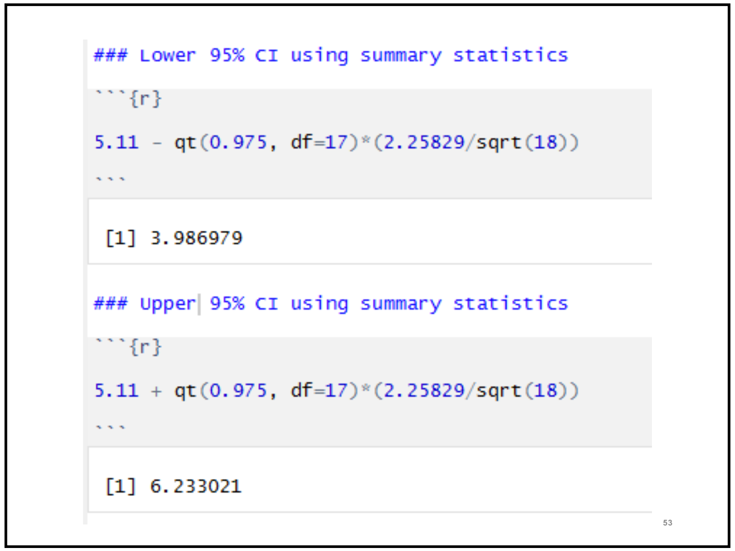
- 95% CI for μ

$$\bar{x} \pm t_{\left(1-\frac{\alpha}{2},n-1\right)} \frac{s}{\sqrt{n}}$$

```
Variable                  N     Mean    StDev
VO2 Improvement          18  5.11111  2.25829
```

```
> qt(0.975, df=17)
[1] 2.109816
```

52



Lower 95% CI using summary statistics

~~~
```{r}
5.11 - qt(0.975, df=17)*(2.25829/sqrt(18))
```
~~~

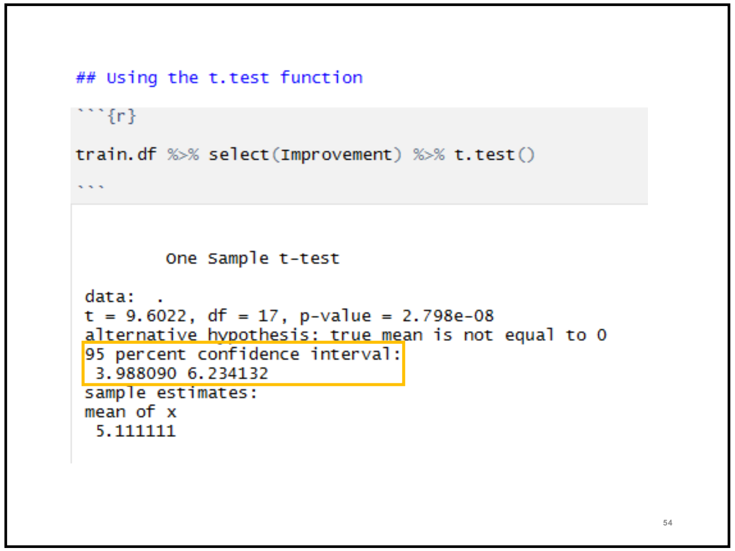
```
[1] 3.986979
```

Upper 95% CI using summary statistics

~~~
```{r}
5.11 + qt(0.975, df=17)*(2.25829/sqrt(18))
```
~~~

```
[1] 6.233021
```

53



Using the t.test function

~~~
```{r}
train.df %>% select(Improvement) %>% t.test()
```
~~~

```
        One Sample t-test

data:  .
t = 9.6022, df = 17, p-value = 2.798e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.988090 6.234132
sample estimates:
mean of x
 5.111111
```

54

Conclusion ?

- On average ?
- What does 95% Confidence mean ?
- Terms and conditions ?
- Random sample ?
- Small n, normality ??

55

If Normality is questionable

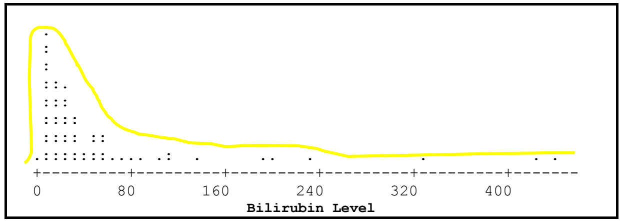
a) Try to transform the data to approximate Normality
- e.g. logarithms or square root

b) Non-Parametric technique
- Bootstrap
- CI for the population MEDIAN

56

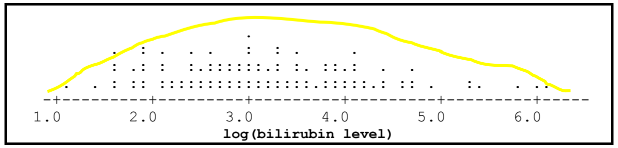
Transforming to Normality

- ***Example:*** A study of Bilirubin levels in patients with Liver Disease



57

Logarithm of Bilirubin Data



1. **Produce an interval estimate for the Population MEAN** ***log*** **bilirubin level**

2. **take anti-logs/exponentials of the resulting interval**

58

If Normality is questionable

a) Try to transform the data to approximate Normality
   - e.g. logarithms or square root

b) Non-Parametric technique
   - Bootstrap
   - CI for the population MEDIAN

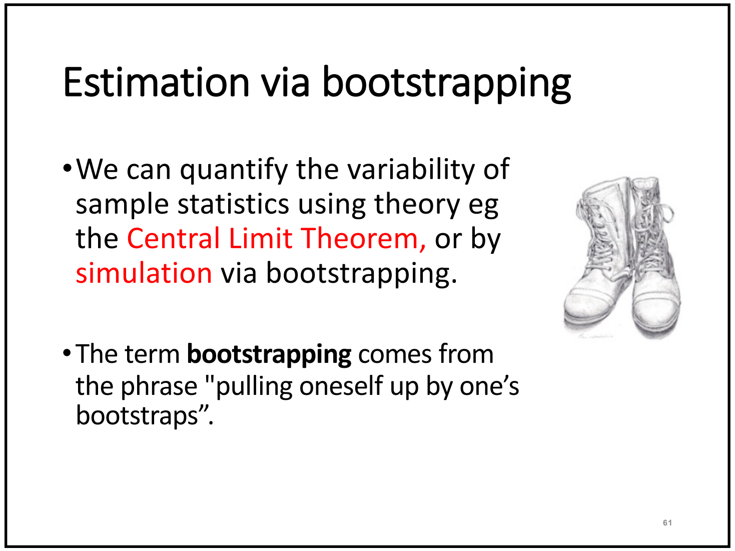
59

The Bootstrap

a) Try to transform the data to approximate Normality
   - e.g. logarithms or square root

b) Non-Parametric technique
   - Bootstrap
   - CI for the population MEDIAN

60

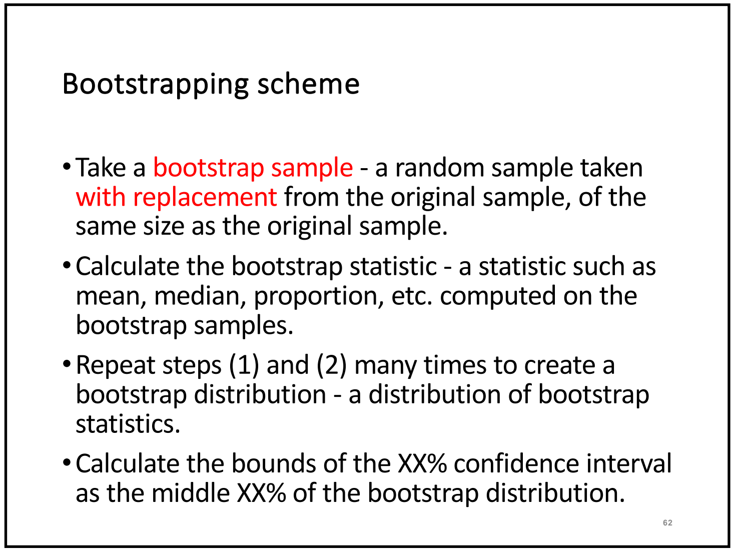


Estimation via bootstrapping

- We can quantify the variability of sample statistics using theory eg the Central Limit Theorem, or by simulation via bootstrapping.
- The term **bootstrapping** comes from the phrase "pulling oneself up by one's bootstraps".

61

61

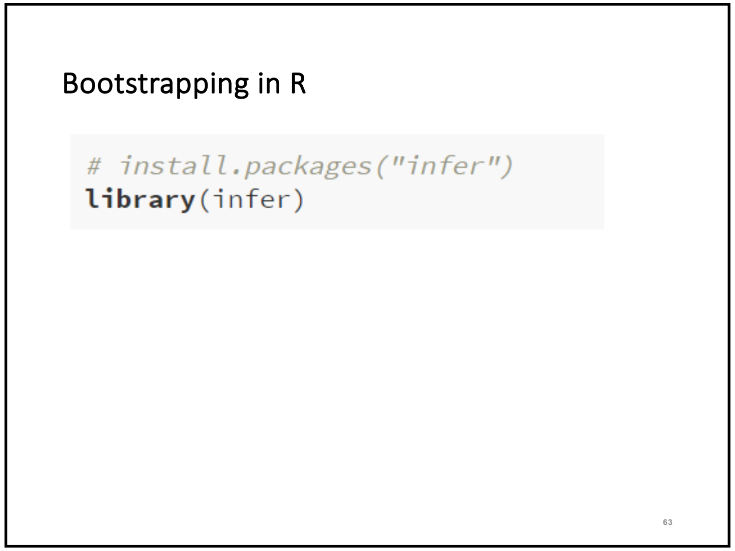


Bootstrapping scheme

- Take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample.
- Calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples.
- Repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics.
- Calculate the bounds of the XX% confidence interval as the middle XX% of the bootstrap distribution.

62

62

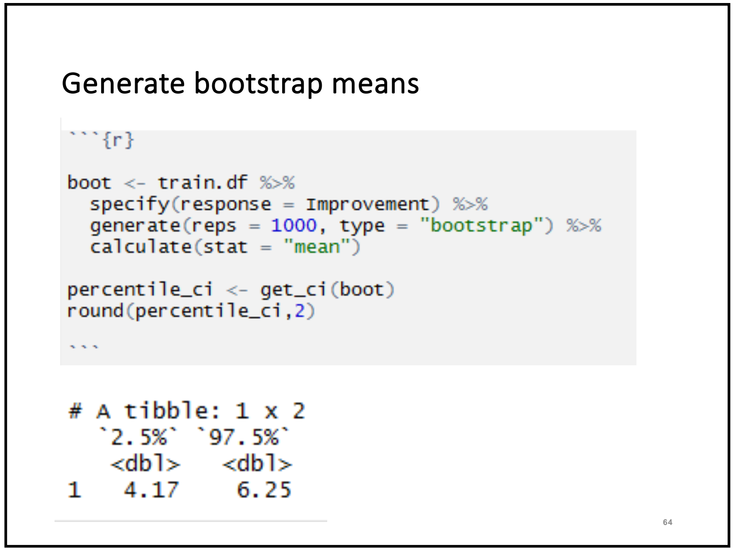


Bootstrapping in R

```
# install.packages("infer")
library(infer)
```

63

63



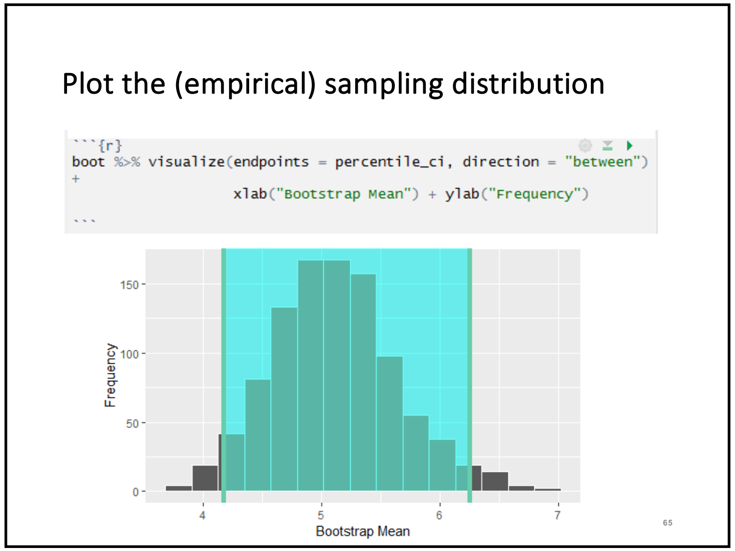
Generate bootstrap means

~~~
```{r}

boot <- train.df %>%
  specify(response = Improvement) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

percentile_ci <- get_ci(boot)
round(percentile_ci,2)

```
~~~

```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1   4.17    6.25
```

64

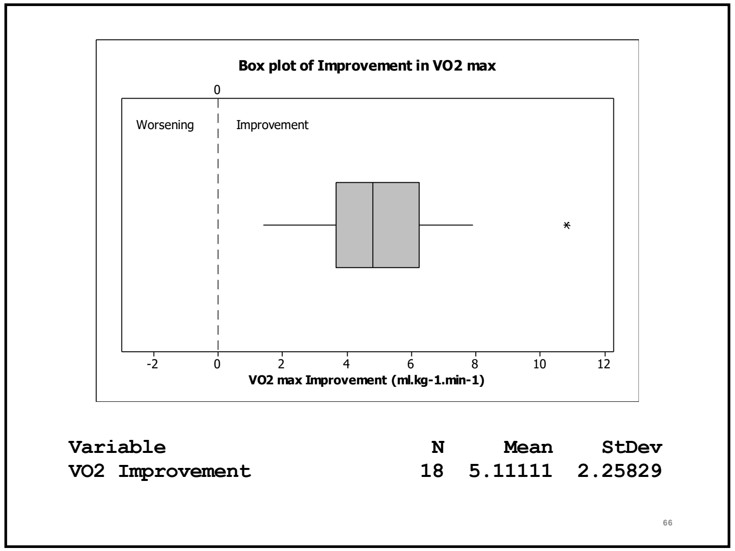
64



Plot the (empirical) sampling distribution

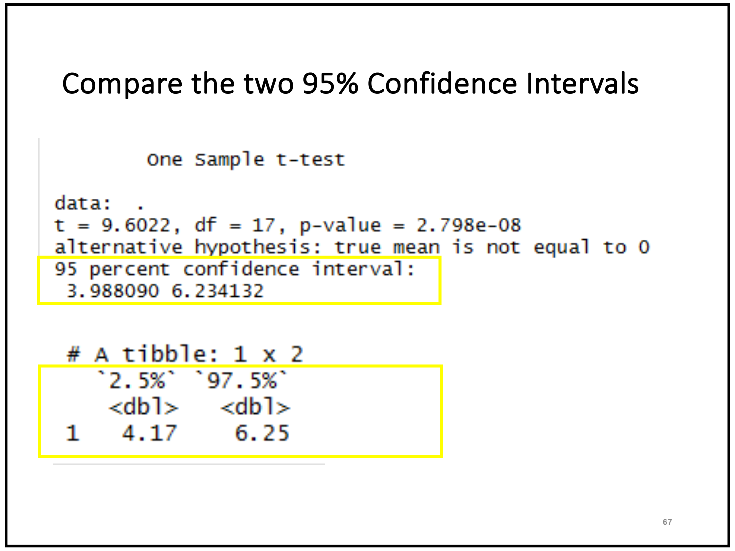
~~~
```{r}
boot %>% visualize(endpoints = percentile_ci, direction = "between")
+
                xlab("Bootstrap Mean") + ylab("Frequency")
```
~~~

65



```
Variable                      N     Mean     StDev
VO2 Improvement              18  5.11111   2.25829
```

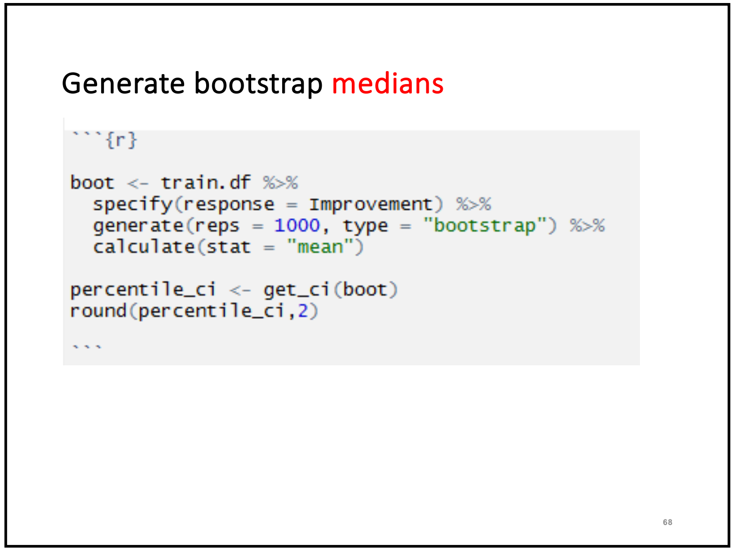
66



Compare the two 95% Confidence Intervals

```
        One Sample t-test

data:  .
t = 9.6022, df = 17, p-value = 2.798e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.988090 6.234132

 # A tibble: 1 x 2
   `2.5%` `97.5%`
    <dbl>    <dbl>
 1   4.17     6.25
```

67

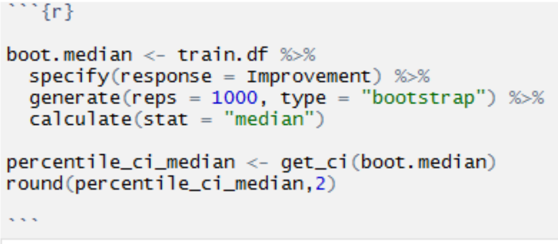


Generate bootstrap medians

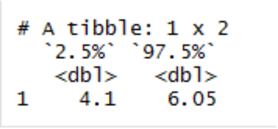
~~~
```{r}

boot <- train.df %>%
  specify(response = Improvement) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

percentile_ci <- get_ci(boot)
round(percentile_ci,2)

```
~~~

68

Generate bootstrap medians



~~~
```{r}

boot.median <- train.df %>%
  specify(response = Improvement) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "median")

percentile_ci_median <- get_ci(boot.median)
round(percentile_ci_median,2)

```
~~~



```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1    4.1    6.05
```

69

69

Celtic Study

- Based on the data provided the sample mean improvement was 5.11 mL/kg/min. We are 95% confident that the typical improvement in VO2 max is likely to be between 4 and 6 mL/kg/min.

- Given that the typical VO2 max at the start of this study was 67.66, the estimated typical improvement is approximately 7% (i.e. 5.11/67.66 expressed as percentage is 0.07*100).

- How would you translate this ?

70

70

Celtic Study

- Does this mean that each player will improve by 5.11 units ?

71

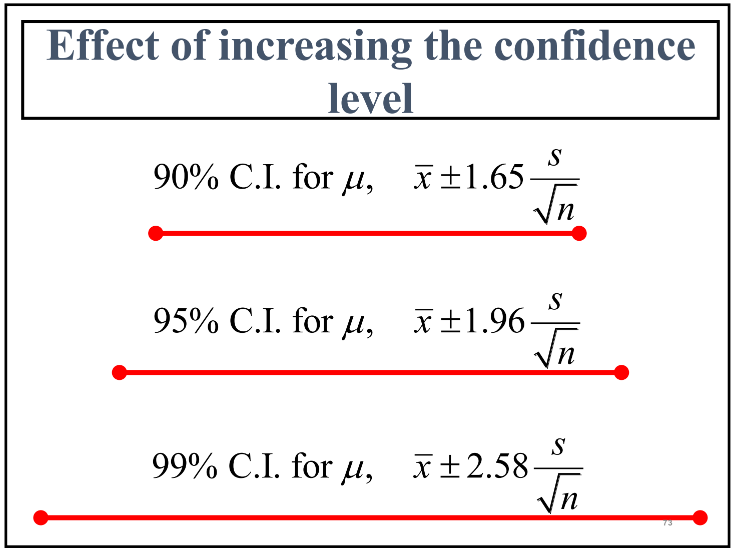
71

Pick a parameter of interest ….

1. *Estimate it using an (unbiased) estimator*
2. *Calculate its corresponding standard error*;
3. *Calculate the corresponding (1-α)100% CI*;
4. Check the terms and conditions
5. Report the conclusions of the analysis.

72

72

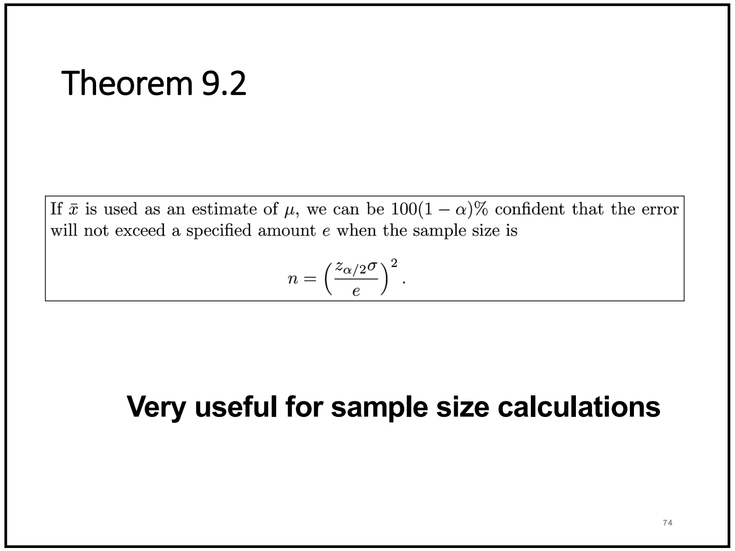


Effect of increasing the confidence level

90% C.I. for μ, $\quad \overline{x} \pm 1.65 \dfrac{s}{\sqrt{n}}$

95% C.I. for μ, $\quad \overline{x} \pm 1.96 \dfrac{s}{\sqrt{n}}$

99% C.I. for μ, $\quad \overline{x} \pm 2.58 \dfrac{s}{\sqrt{n}}$



Theorem 9.2

If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Very useful for sample size calculations

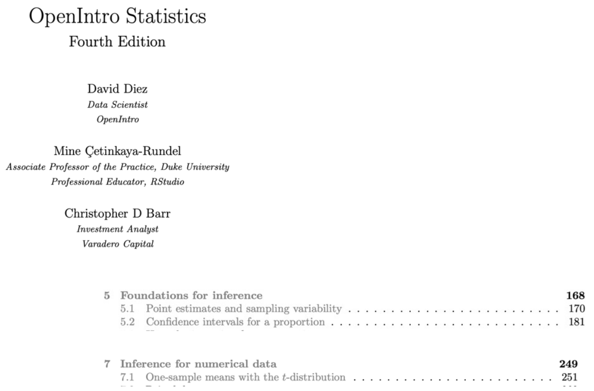
Topic 10: Hypothesis testing

1

Learning Outcomes

1. Carry out hypothesis tests for a single mean.
2. Use the *p*-value approach for making decisions in hypothesis tests.
3. Understand types of testing errors
4. Understand the relationship between hypothesis testing and confidence intervals and the advantages of interval estimation
5. Additional reading material : Open Intro book Chapters 5.1 & 7.1

2



OpenIntro Statistics
Fourth Edition

David Diez
Data Scientist
OpenIntro

Mine Çetinkaya-Rundel
Associate Professor of the Practice, Duke University
Professional Educator, RStudio

Christopher D Barr
Investment Analyst
Varadero Capital

3

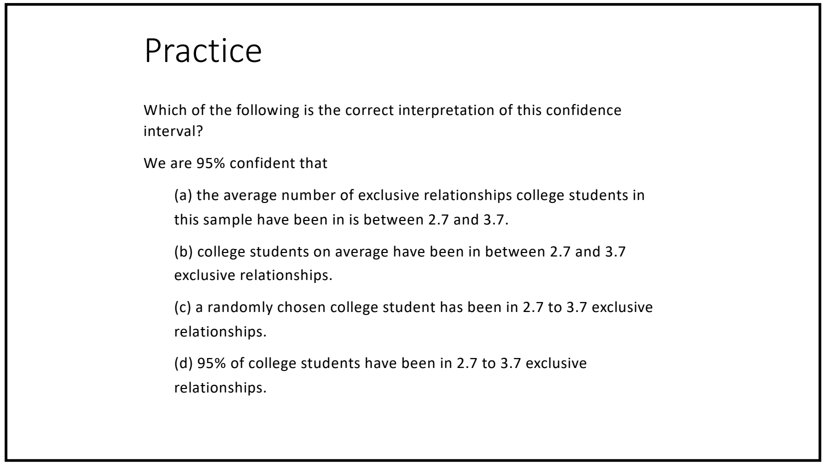
Recap: Inference using Confidence Interval Estimation

A claim has been made that college students have been in, on average, at least 4 exclusive relationships. Data collected on a random sample of 50 college students yielded a mean of 3.2 and a standard deviation of 1.74.

Do these data provide evidence for or against the hypothesis claimed ?

The corresponding 95% CI is [2.7, 3.7].

4



Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

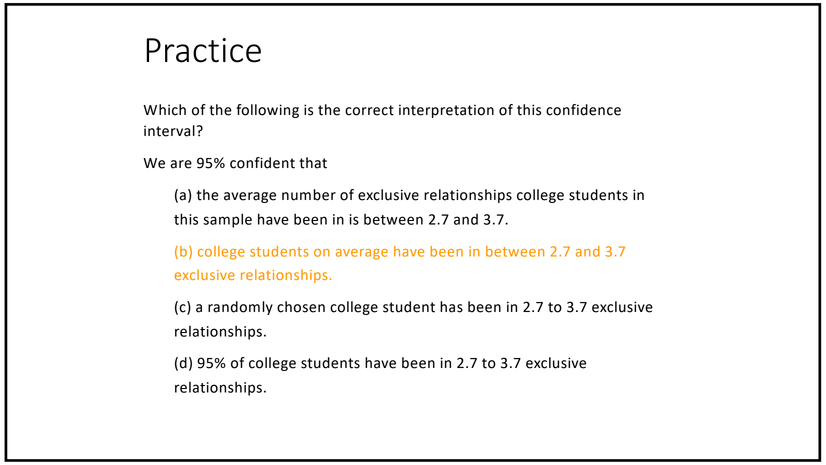
(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

5



Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

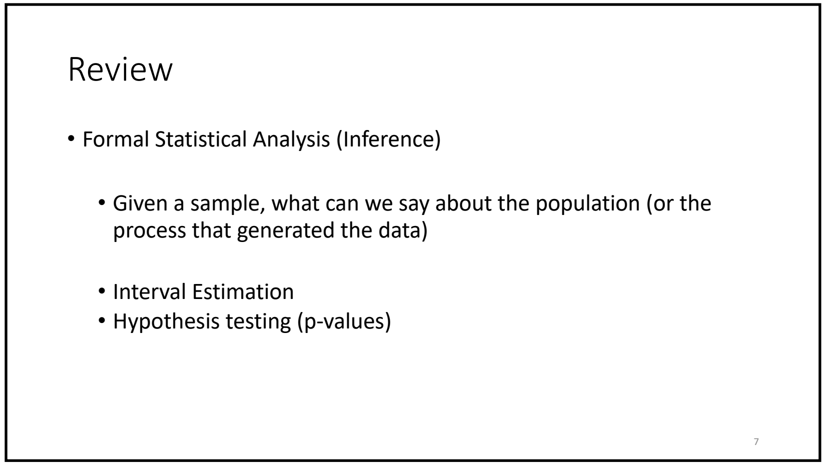
(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

6

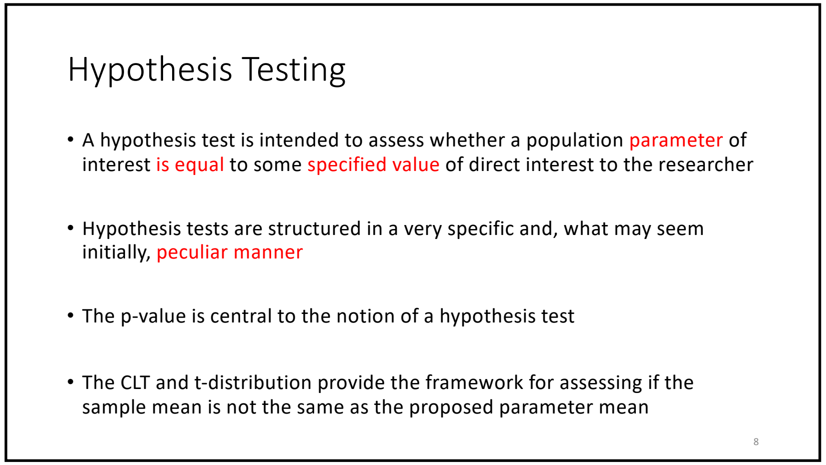


Review

- Formal Statistical Analysis (Inference)
  - Given a sample, what can we say about the population (or the process that generated the data)
  - Interval Estimation
  - Hypothesis testing (p-values)

7

7



Hypothesis Testing

- A hypothesis test is intended to assess whether a population parameter of interest is equal to some specified value of direct interest to the researcher
- Hypothesis tests are structured in a very specific and, what may seem initially, peculiar manner
- The p-value is central to the notion of a hypothesis test
- The CLT and t-distribution provide the framework for assessing if the sample mean is not the same as the proposed parameter mean

8

8

Null and alternative hypotheses

- The null hypothesis is a claim to be tested – often the skeptical claim of "no effect".. eg

$$H_0:\ \mu = \mu_0$$

- The alternative hypothesis is an alternative claim under consideration, often represented by a range of parameter values – eg

$$H_1:\ \mu \neq \mu_0$$

- We only reject the null in favour of the alternative if there is strong supporting evidence.
- We decide a priori how much evidence is "strong" enough to reject the null

9

9

Stages in Hypothesis Testing

1. Null Hypothesis: The hypothesis that the population parameter is equal to some claimed value (H_0)
2. Study or Alternative Hypothesis: The hypothesis that must be true if the null hypothesis is false (H_1)
3. **Collect appropriate data**
4. Assess, through a test statistic, how probable (the p-value) it would be to observe data as or more extreme than the data actually collected if, in fact, the Null Hypothesis was true
5. Come to a conclusion whether or not to reject the Null Hypothesis

10

10

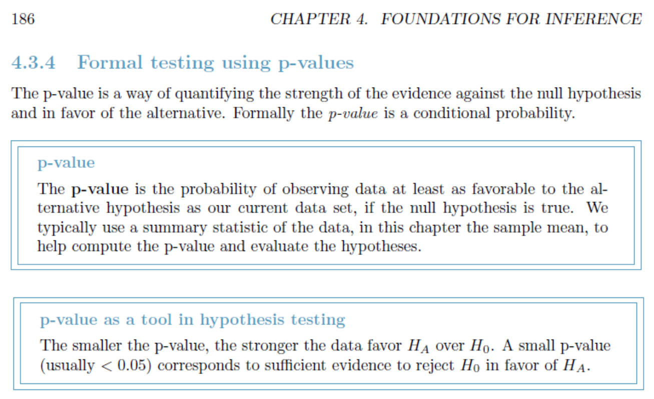
Rejecting/not rejecting the null

- If we do not reject the null hypothesis in favour of the alternative, we are saying that the effect indicated by the sample is due only to sampling variation.

- If we do reject the null hypothesis in favour of the alternative, we are saying that the effect indicated by the sample is real, in that it is more than can be attributed to sampling variation.

11

11



186 CHAPTER 4. FOUNDATIONS FOR INFERENCE

4.3.4 Formal testing using p-values

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative. Formally the *p-value* is a conditional probability.

p-value

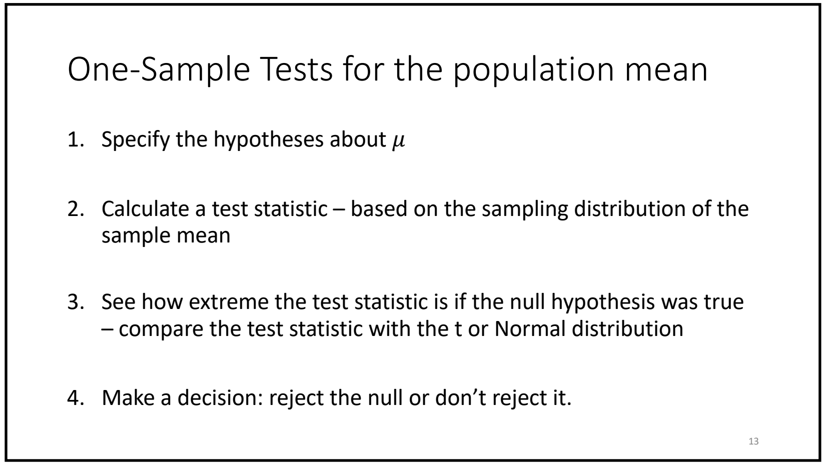
The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true. We typically use a summary statistic of the data, in this chapter the sample mean, to help compute the p-value and evaluate the hypotheses.

p-value as a tool in hypothesis testing

The smaller the p-value, the stronger the data favor H_A over H_0. A small p-value (usually < 0.05) corresponds to sufficient evidence to reject H_0 in favor of H_A.

12

12

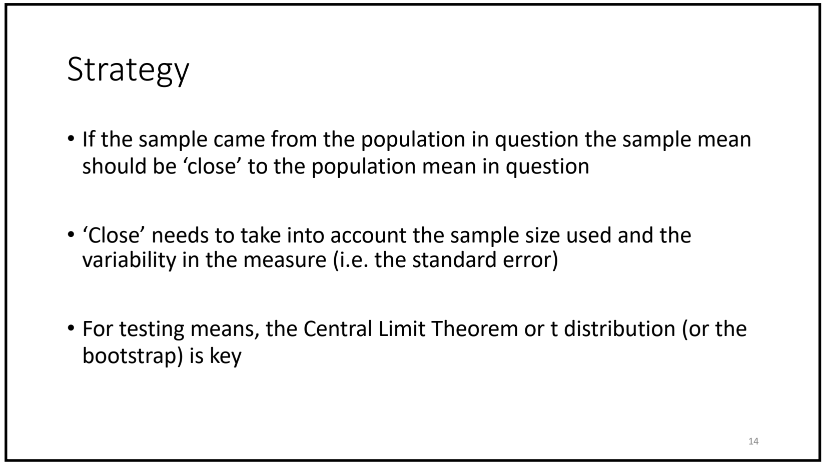


One-Sample Tests for the population mean

1. Specify the hypotheses about μ
2. Calculate a test statistic – based on the sampling distribution of the sample mean
3. See how extreme the test statistic is if the null hypothesis was true – compare the test statistic with the t or Normal distribution
4. Make a decision: reject the null or don't reject it.

13

13

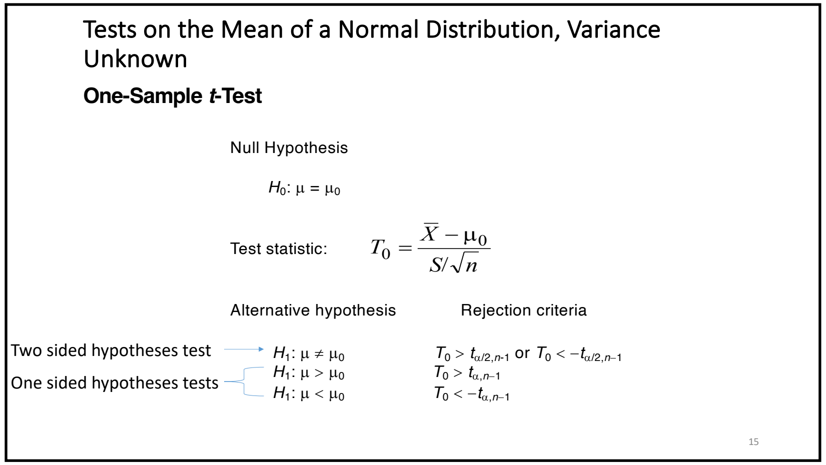


Strategy

- If the sample came from the population in question the sample mean should be 'close' to the population mean in question
- 'Close' needs to take into account the sample size used and the variability in the measure (i.e. the standard error)
- For testing means, the Central Limit Theorem or t distribution (or the bootstrap) is key

14

14



Tests on the Mean of a Normal Distribution, Variance Unknown

One-Sample *t*-Test

Null Hypothesis

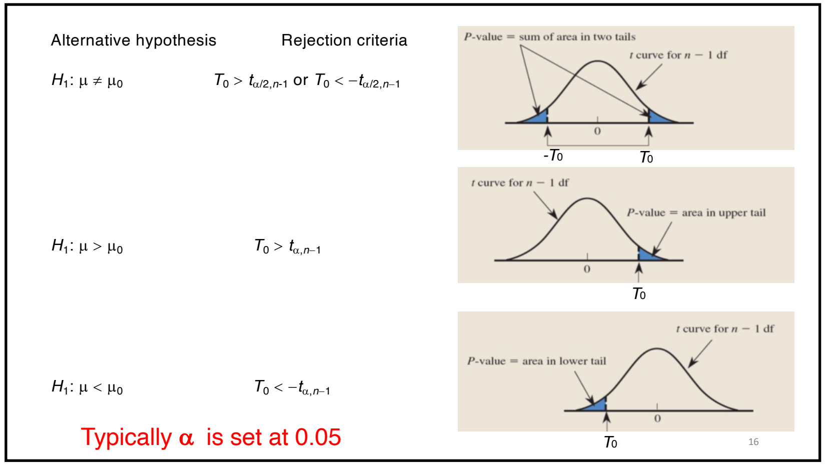
H_0: $\mu = \mu_0$

Test statistic: $T_0 = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$

	Alternative hypothesis	Rejection criteria
Two sided hypotheses test →	H_1: $\mu \neq \mu_0$	$T_0 > t_{\alpha/2,n-1}$ or $T_0 < -t_{\alpha/2,n-1}$
One sided hypotheses tests	H_1: $\mu > \mu_0$	$T_0 > t_{\alpha,n-1}$
	H_1: $\mu < \mu_0$	$T_0 < -t_{\alpha,n-1}$

15

15

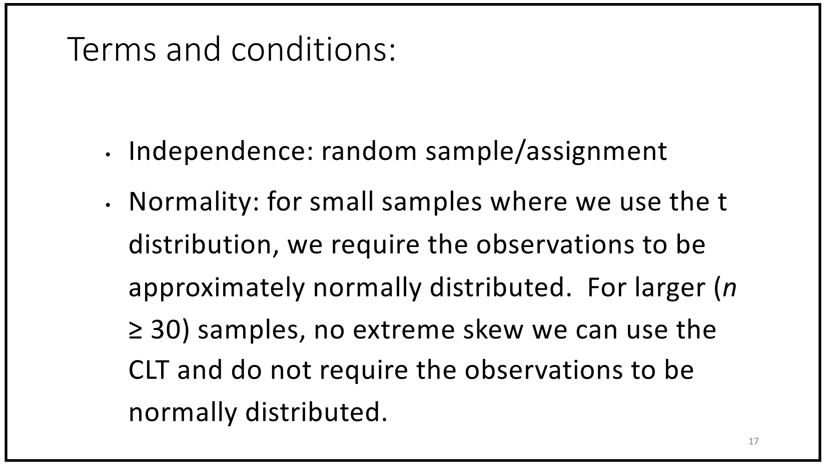


Alternative hypothesis	Rejection criteria
H_1: $\mu \neq \mu_0$	$T_0 > t_{\alpha/2,n-1}$ or $T_0 < -t_{\alpha/2,n-1}$
H_1: $\mu > \mu_0$	$T_0 > t_{\alpha,n-1}$
H_1: $\mu < \mu_0$	$T_0 < -t_{\alpha,n-1}$

Typically α is set at 0.05

16

16

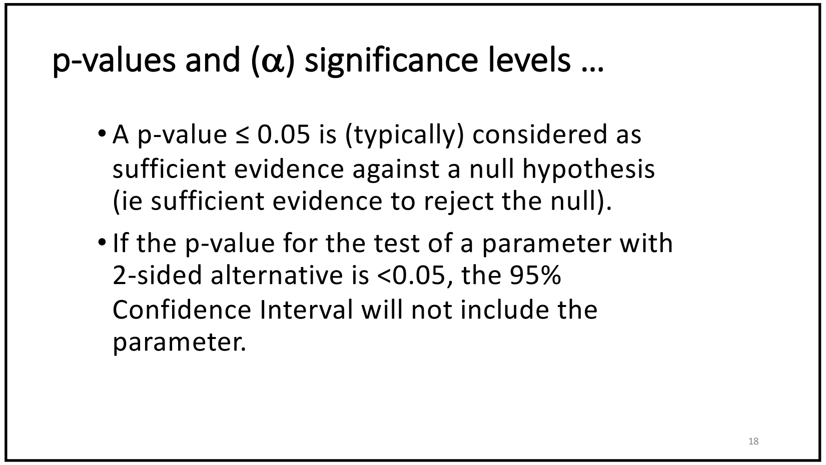


Terms and conditions:

- Independence: random sample/assignment
- Normality: for small samples where we use the t distribution, we require the observations to be approximately normally distributed. For larger ($n \geq 30$) samples, no extreme skew we can use the CLT and do not require the observations to be normally distributed.

17

17

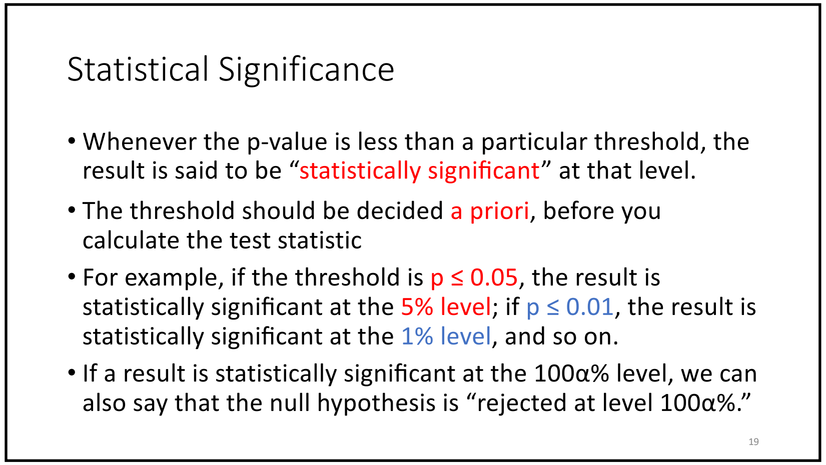


p-values and (α) significance levels …

- A p-value ≤ 0.05 is (typically) considered as sufficient evidence against a null hypothesis (ie sufficient evidence to reject the null).
- If the p-value for the test of a parameter with 2-sided alternative is <0.05, the 95% Confidence Interval will not include the parameter.

18

18

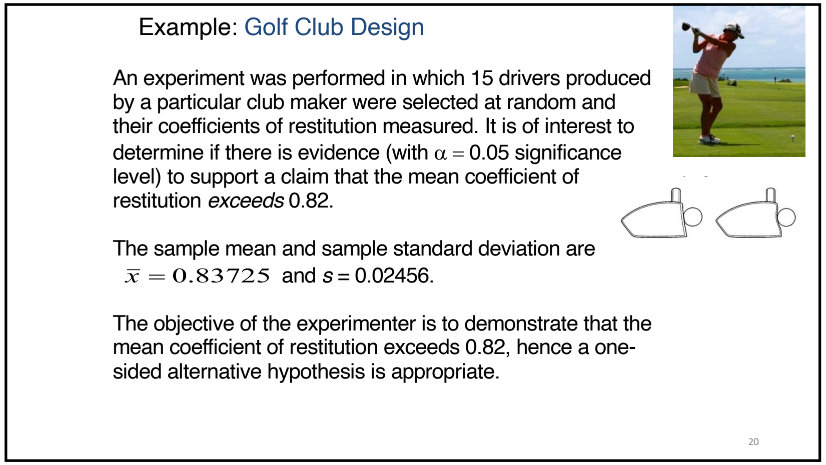


Statistical Significance

- Whenever the p-value is less than a particular threshold, the result is said to be "statistically significant" at that level.
- The threshold should be decided a priori, before you calculate the test statistic
- For example, if the threshold is $p \leq 0.05$, the result is statistically significant at the 5% level; if $p \leq 0.01$, the result is statistically significant at the 1% level, and so on.
- If a result is statistically significant at the 100α% level, we can also say that the null hypothesis is "rejected at level 100α%."

19

19



Example: Golf Club Design

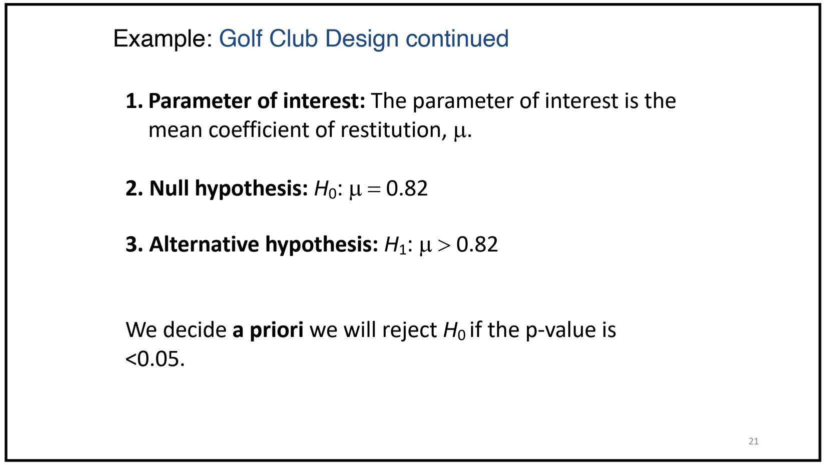
An experiment was performed in which 15 drivers produced by a particular club maker were selected at random and their coefficients of restitution measured. It is of interest to determine if there is evidence (with $\alpha = 0.05$ significance level) to support a claim that the mean coefficient of restitution *exceeds* 0.82.

The sample mean and sample standard deviation are $\bar{x} = 0.83725$ and $s = 0.02456$.

The objective of the experimenter is to demonstrate that the mean coefficient of restitution exceeds 0.82, hence a one-sided alternative hypothesis is appropriate.

20

20



Example: Golf Club Design continued

1. Parameter of interest: The parameter of interest is the mean coefficient of restitution, μ.

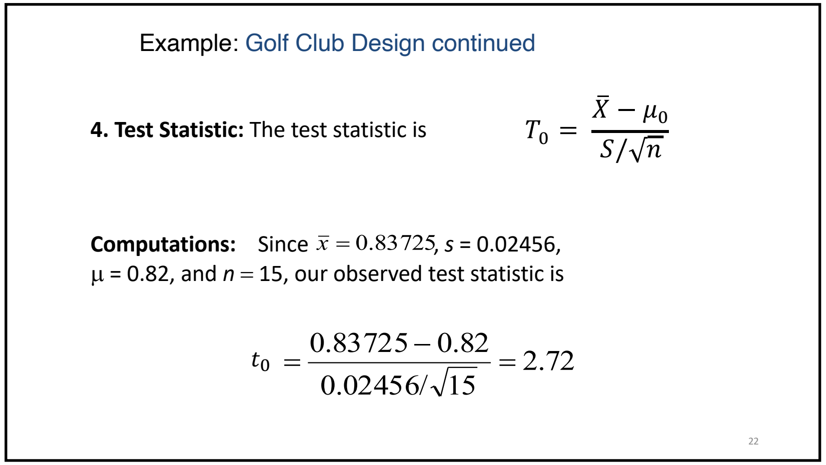
2. Null hypothesis: H_0: $\mu = 0.82$

3. Alternative hypothesis: H_1: $\mu > 0.82$

We decide **a priori** we will reject H_0 if the p-value is <0.05.

21

21



Example: Golf Club Design continued

4. Test Statistic: The test statistic is $T_0 = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$

Computations: Since $\bar{x} = 0.83725$, $s = 0.02456$, $\mu = 0.82$, and $n = 15$, our observed test statistic is

$$t_0 = \frac{0.83725 - 0.82}{0.02456/\sqrt{15}} = 2.72$$

22

22

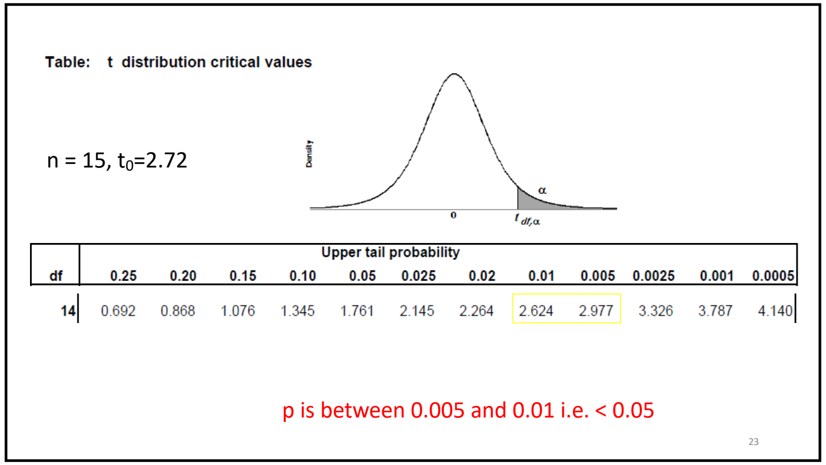


Table: t distribution critical values

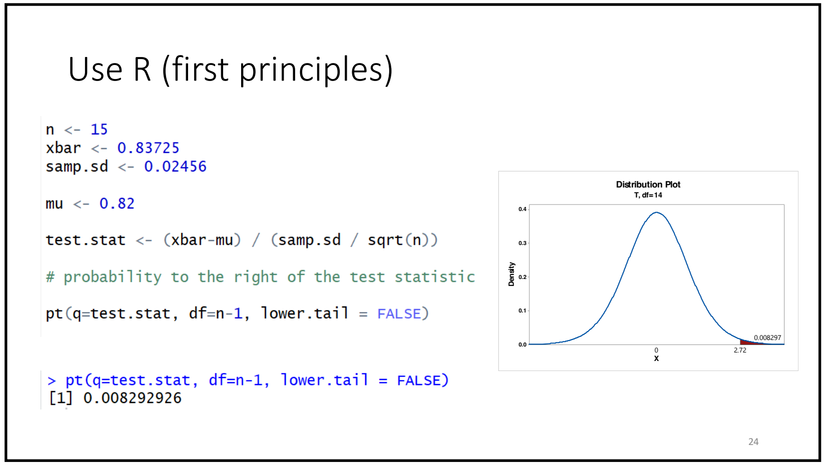
n = 15, t_0=2.72

	Upper tail probability											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140

p is between 0.005 and 0.01 i.e. < 0.05

23

23



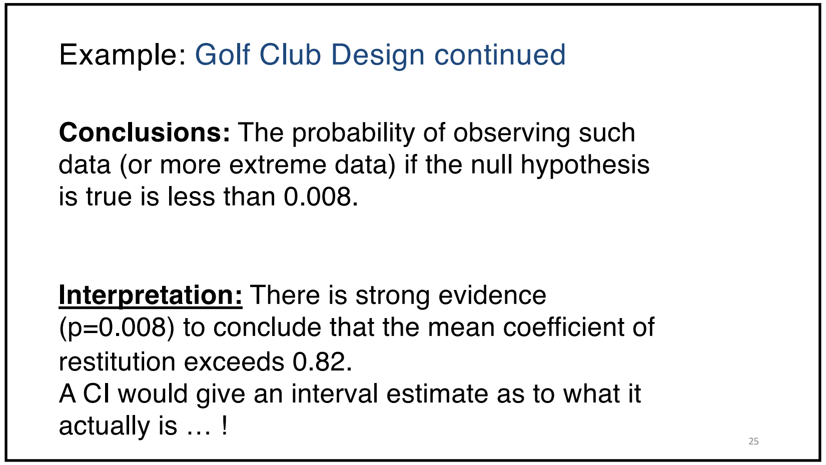
Use R (first principles)

```
n <- 15
xbar <- 0.83725
samp.sd <- 0.02456

mu <- 0.82

test.stat <- (xbar-mu) / (samp.sd / sqrt(n))

# probability to the right of the test statistic

pt(q=test.stat, df=n-1, lower.tail = FALSE)
```

```
> pt(q=test.stat, df=n-1, lower.tail = FALSE)
[1] 0.008292926
```

24

24



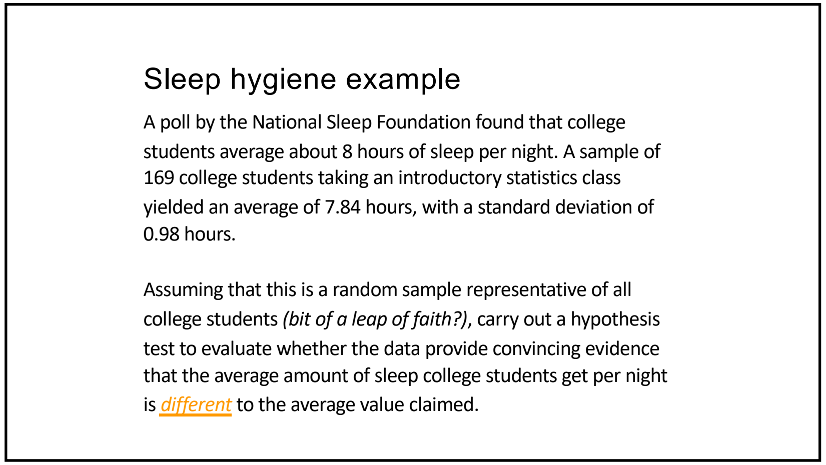
Example: Golf Club Design continued

Conclusions: The probability of observing such data (or more extreme data) if the null hypothesis is true is less than 0.008.

Interpretation: There is strong evidence (p=0.008) to conclude that the mean coefficient of restitution exceeds 0.82.
A CI would give an interval estimate as to what it actually is … !

25

25

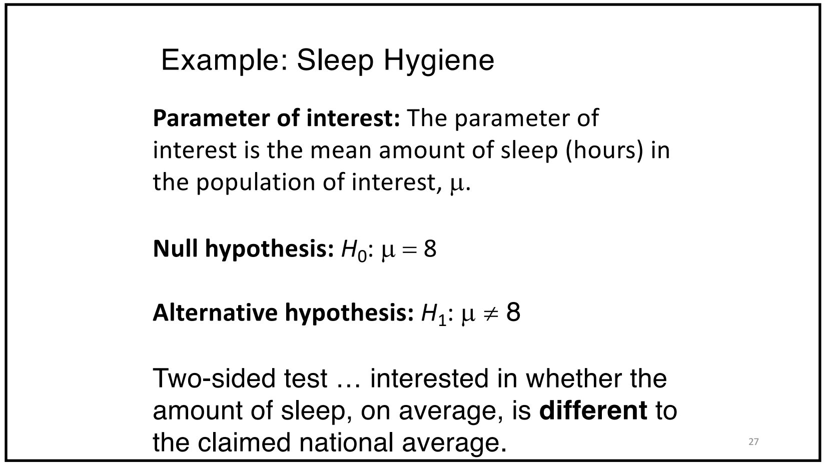


Sleep hygiene example

A poll by the National Sleep Foundation found that college students average about 8 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 7.84 hours, with a standard deviation of 0.98 hours.

Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, carry out a hypothesis test to evaluate whether the data provide convincing evidence that the average amount of sleep college students get per night is *different* to the average value claimed.

26



Example: Sleep Hygiene

Parameter of interest: The parameter of interest is the mean amount of sleep (hours) in the population of interest, μ.

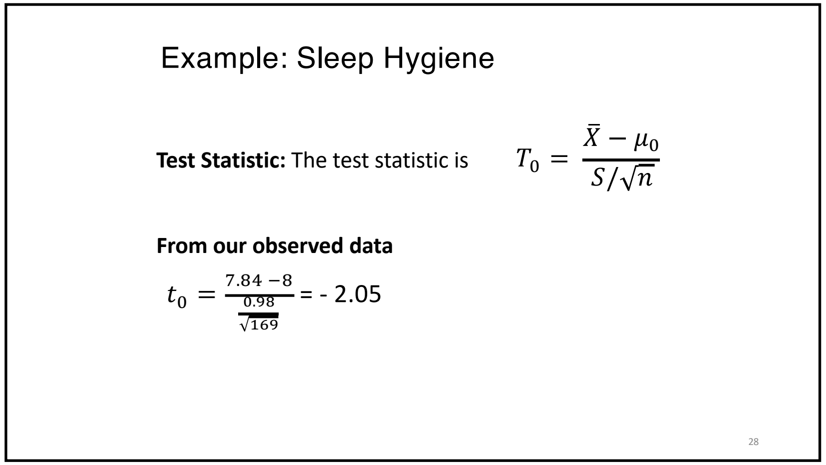
Null hypothesis: H_0: $\mu = 8$

Alternative hypothesis: H_1: $\mu \neq 8$

Two-sided test … interested in whether the amount of sleep, on average, is **different** to the claimed national average.

27

27



Example: Sleep Hygiene

Test Statistic: The test statistic is $T_0 = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$

From our observed data

$$t_0 = \frac{7.84 - 8}{\frac{0.98}{\sqrt{169}}} = -\,2.05$$

28

28

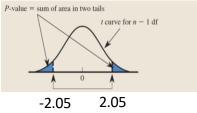
Example: Sleep Hygiene

Test Statistic: The test statistic is $T_0 = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$

From our observed data

$$t_0 = \frac{7.84 - 8}{\frac{0.98}{\sqrt{169}}} = -\ 2.05$$

p-value : calculate area to the right of 2.05 and to the left of -2.05 in a t distribution with 169-1 degrees of freedom.

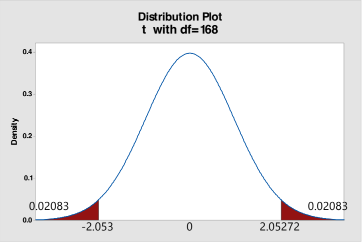


29

29

Example: Sleep Hygiene

P-value : calculate area to the right of 2.05 and to the left of -2.05 in a t distribution with 169-1 degrees of freedom.



30

30

Example: Sleep Hygiene

p-value : use the symmetry in the distribution i.e. calculate area to the right of 2.05 in a t distribution with 169-1 degrees of freedom and double it.

```
> 2 * pt(q= 2.052717, df=168, lower.tail = FALSE)
[1] 0.04165098
```

p-value : 0.0416

31

31

Example: Sleep Hygiene

Conclusions: The probability of observing such data (or more extreme) if the null hypothesis is true is = 0.04.

<u>Interpretation:</u> As the p-value is less than 0.05, there is evidence (at the 5% significance level) that the mean hours sleeping is different from the national average of 8.

32

32

Example: Sleep Hygiene using R

Parameter of interest: The parameter of interest is the mean amount of sleep (hours) in the population of interest, μ.

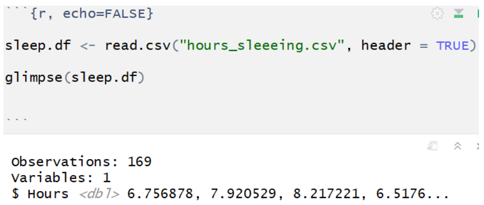
Null hypothesis: H_0: $\mu = 8$

Alternative hypothesis: H_1: $\mu \neq 8$

Two-sided test … interested in whether the amount of sleep, on average, is different to claimed national average.

33

33

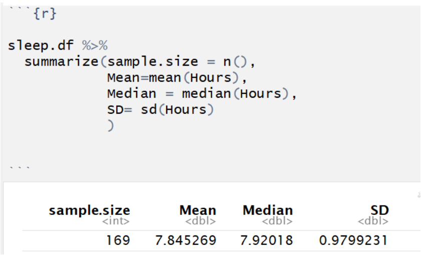


~~~
```{r, echo=FALSE}
sleep.df <- read.csv("hours_sleeeing.csv", header = TRUE)
glimpse(sleep.df)

```
~~~

```
Observations: 169
Variables: 1
$ Hours <dbl> 6.756878, 7.920529, 8.217221, 6.5176...
```

34

34

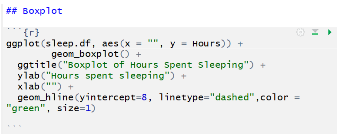


~~~
```{r}
sleep.df %>%
  summarize(sample.size = n(),
            Mean=mean(Hours),
            Median = median(Hours),
            SD= sd(Hours)
            )

```
~~~

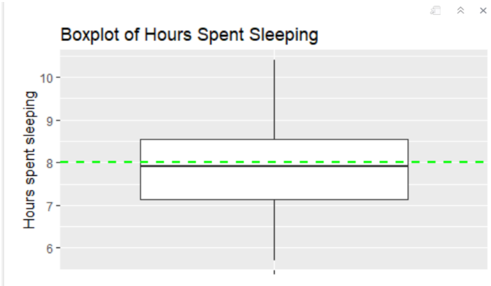
sample.size <int>	Mean <dbl>	Median <dbl>	SD <dbl>
169	7.845269	7.92018	0.9799231

35

35

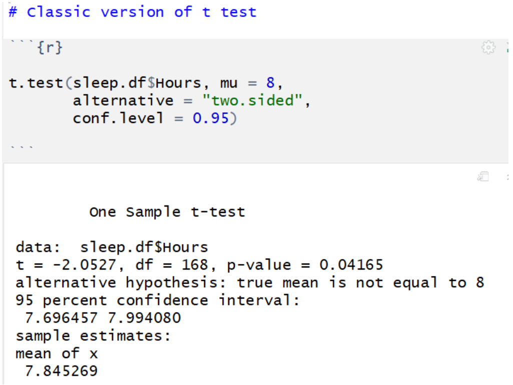


~~~
## Boxplot

```{r}
ggplot(sleep.df, aes(x = "", y = Hours)) +
        geom_boxplot() +
  ggtitle("Boxplot of Hours Spent Sleeping") +
  ylab("Hours spent sleeping") +
  xlab("") +
  geom_hline(yintercept=8, linetype="dashed",color =
"green", size=1)

```
~~~



36

36



~~~
# Classic version of t test
```{r}

t.test(sleep.df$Hours, mu = 8,
       alternative = "two.sided",
       conf.level = 0.95)

```


        One Sample t-test

data:  sleep.df$Hours
t = -2.0527, df = 168, p-value = 0.04165
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 7.696457 7.994080
sample estimates:
mean of x
 7.845269
~~~

37

Statistical Significance Is Not the Same as Practical Significance

- When a result has a small p-value, we say that it is “statistically significant.” In common usage, the word significant means “important.” It is therefore tempting to think that statistically significant results must always be important.

- This is not the case. Sometimes statistically significant results do not have any scientific or practical importance.

- A difference is only a difference if it makes a difference.

38

Statistical Significance Is Not the Same as Practical Significance continued …

- The p-value does not measure practical significance. What it does measure is the degree of confidence we can have that the true value is really different from the value specified by the null hypothesis.

- When the p-value is small, then we can be confident that the true value is really different. This does not necessarily imply that the difference is large enough to be of practical importance.

39

Connection between Hypothesis Tests and Confidence Intervals

A close relationship exists between the test of a hypothesis for θ, and the confidence interval for θ.

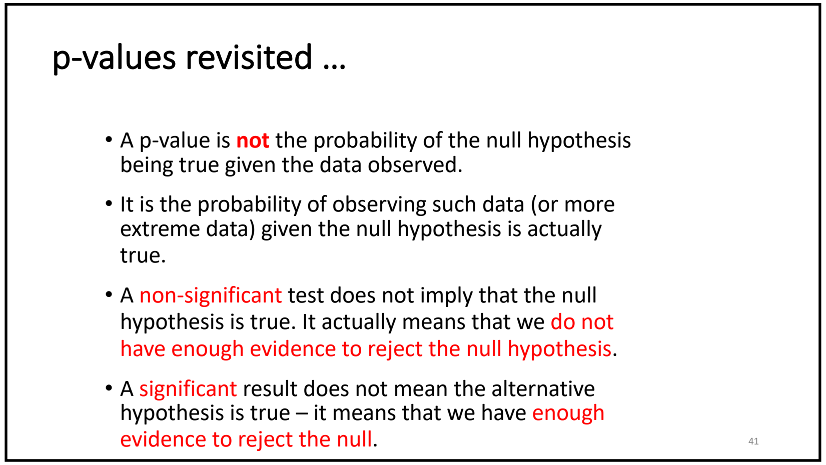
If $[l, u]$ is a 95% confidence interval for the parameter θ, the test of the null hypothesis against a 2-sided alternative at the 0.05 significance level

$$H_0: \theta = \theta_0$$
$$H_1: \theta \neq \theta_0$$

will lead to rejection of H_0 if and only if θ_0 is **not** in the 95% CI $[l, u]$.

And similarly for your alpha of choice e.g. 90% CI and p < 0.10 ….

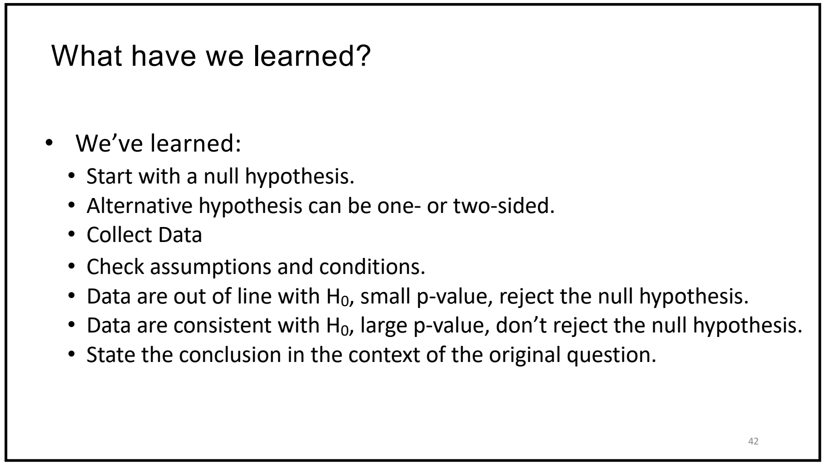
40



p-values revisited …

- A p-value is **not** the probability of the null hypothesis being true given the data observed.
- It is the probability of observing such data (or more extreme data) given the null hypothesis is actually true.
- A non-significant test does not imply that the null hypothesis is true. It actually means that we do not have enough evidence to reject the null hypothesis.
- A significant result does not mean the alternative hypothesis is true – it means that we have enough evidence to reject the null.

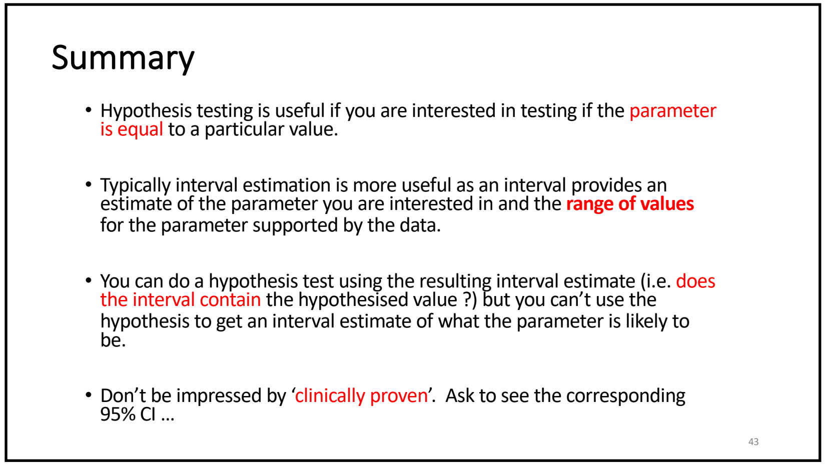
41



What have we learned?

- We've learned:
  - Start with a null hypothesis.
  - Alternative hypothesis can be one- or two-sided.
  - Collect Data
  - Check assumptions and conditions.
  - Data are out of line with H_0, small p-value, reject the null hypothesis.
  - Data are consistent with H_0, large p-value, don't reject the null hypothesis.
  - State the conclusion in the context of the original question.

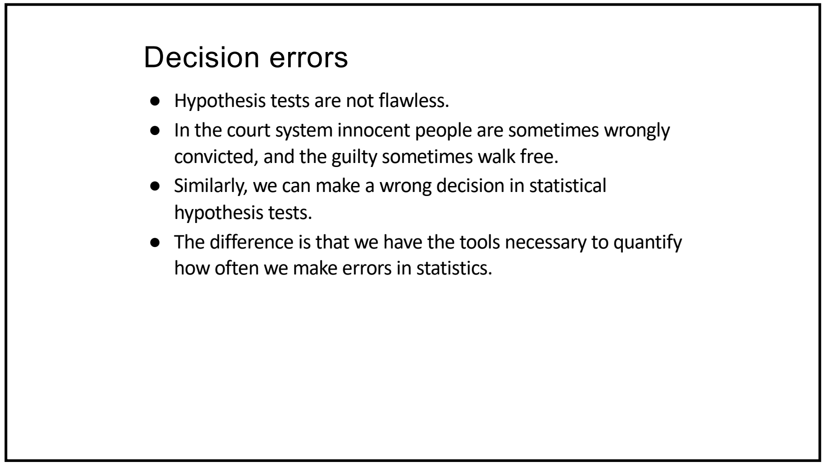
42



Summary

- Hypothesis testing is useful if you are interested in testing if the parameter is equal to a particular value.
- Typically interval estimation is more useful as an interval provides an estimate of the parameter you are interested in and the **range of values** for the parameter supported by the data.
- You can do a hypothesis test using the resulting interval estimate (i.e. does the interval contain the hypothesised value ?) but you can't use the hypothesis to get an interval estimate of what the parameter is likely to be.
- Don't be impressed by 'clinically proven'. Ask to see the corresponding 95% CI …

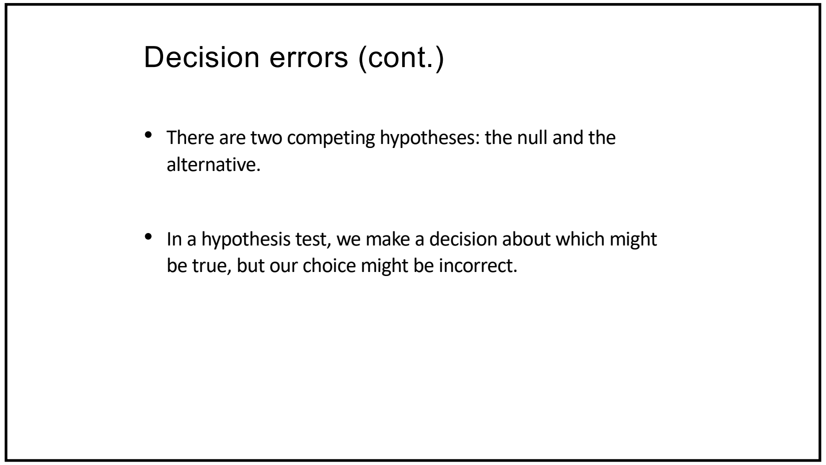
43



Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted, and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

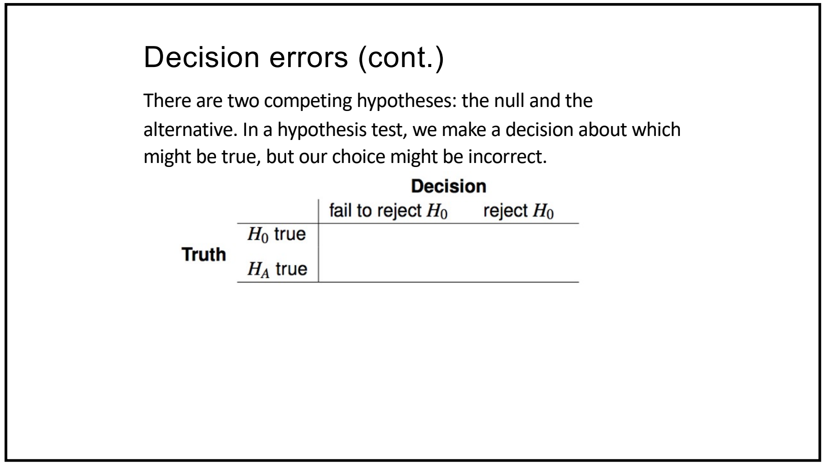
44



Decision errors (cont.)

- There are two competing hypotheses: the null and the alternative.
- In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

45

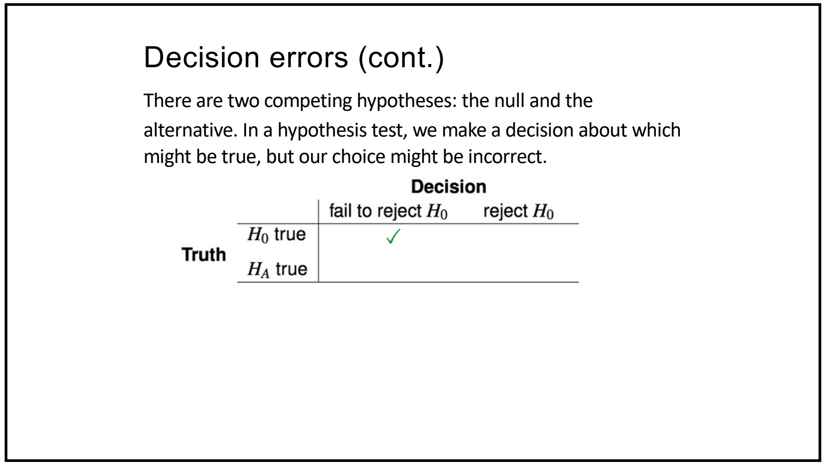


Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

46

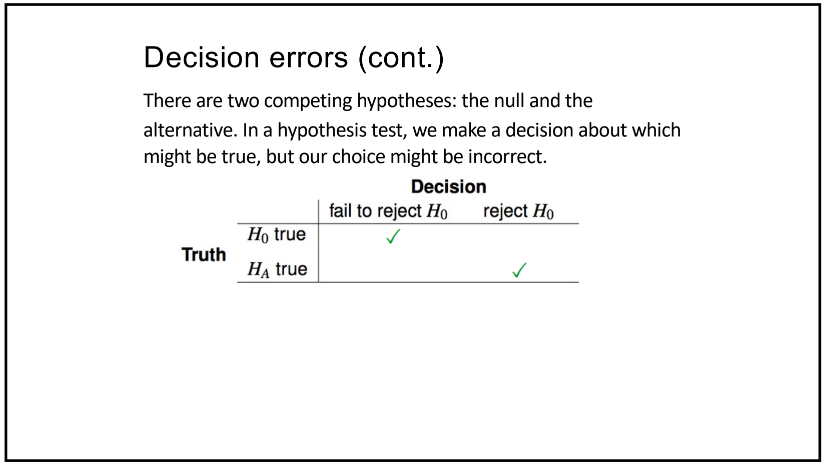


Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		

47

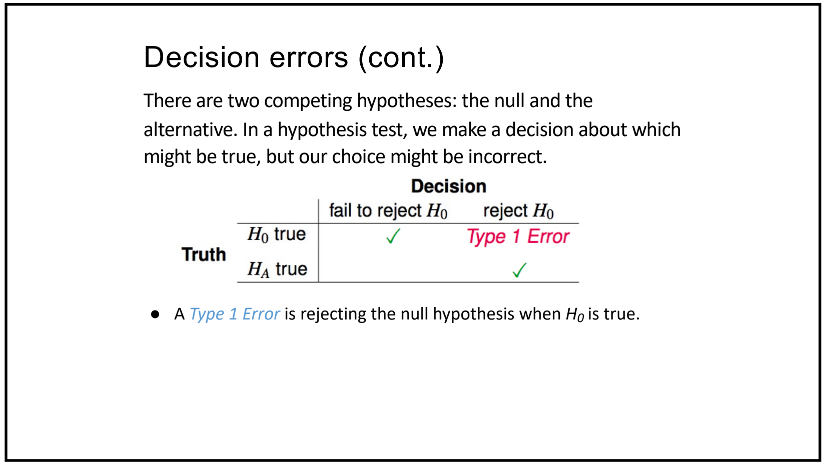


Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		✓

48



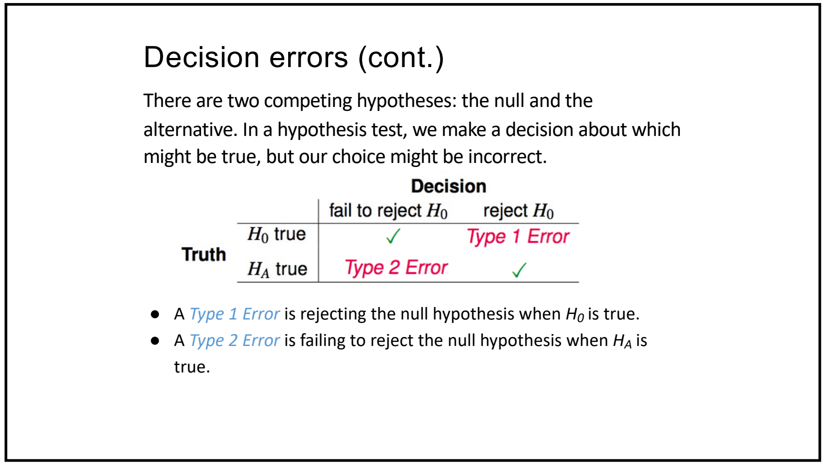
Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true		✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.

49



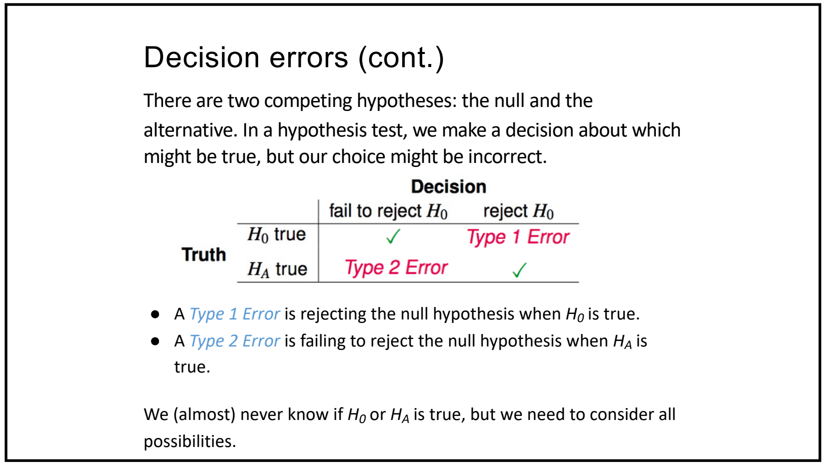
Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.

50



Decision errors (cont.)

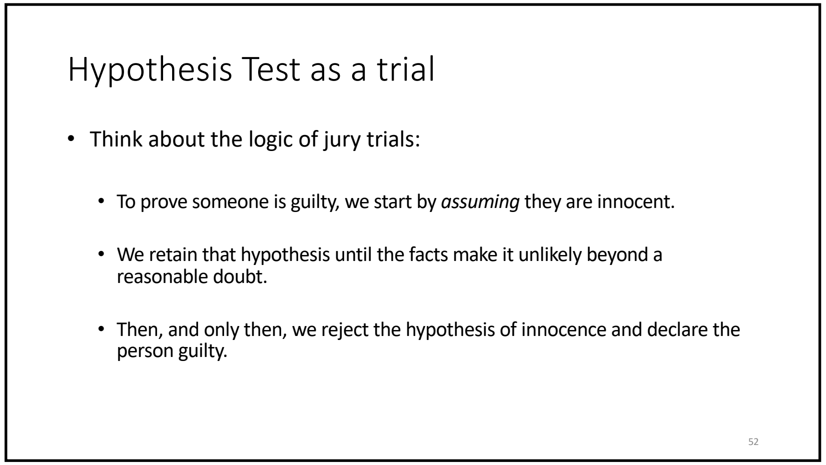
There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.

We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

51

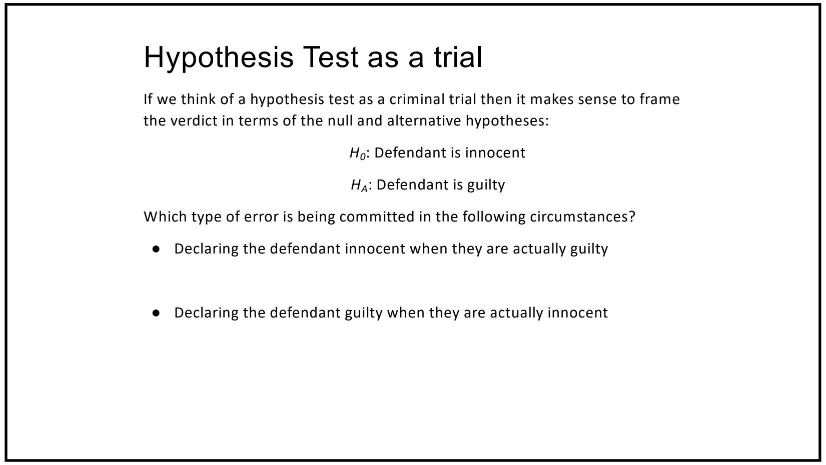


Hypothesis Test as a trial

- Think about the logic of jury trials:
  - To prove someone is guilty, we start by *assuming* they are innocent.
  - We retain that hypothesis until the facts make it unlikely beyond a reasonable doubt.
  - Then, and only then, we reject the hypothesis of innocence and declare the person guilty.

52

52



Hypothesis Test as a trial

If we think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

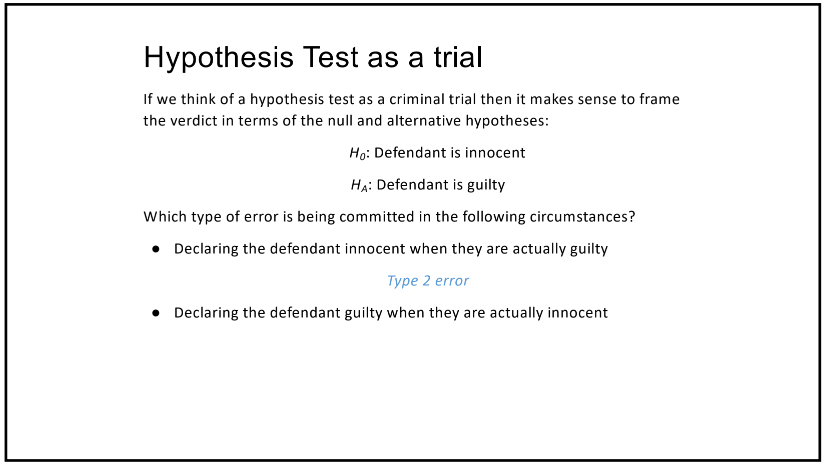
H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty
- Declaring the defendant guilty when they are actually innocent

53



Hypothesis Test as a trial

If we think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

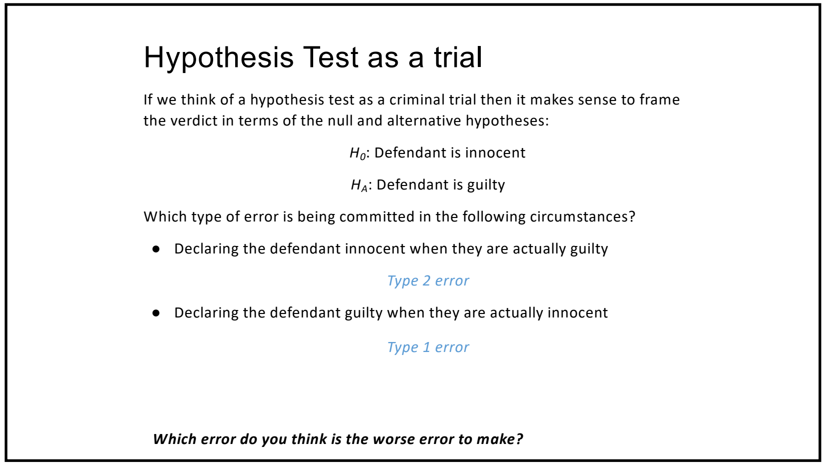
Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

54



Hypothesis Test as a trial

If we think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

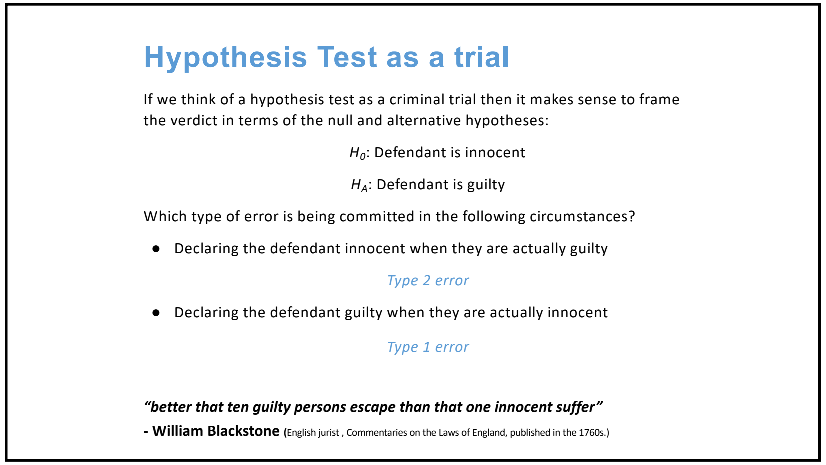
  Type 2 error

- Declaring the defendant guilty when they are actually innocent

  Type 1 error

Which error do you think is the worse error to make?

55



Hypothesis Test as a trial

If we think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

  Type 1 error

"better that ten guilty persons escape than that one innocent suffer"

- William Blackstone (English jurist , Commentaries on the Laws of England, published in the 1760s.)

56

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.

57

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.

58

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error}) = \alpha$$

Or $$P(\text{Reject H0} \mid \text{H0 true}) = \alpha$$

59

Type 1 error rate

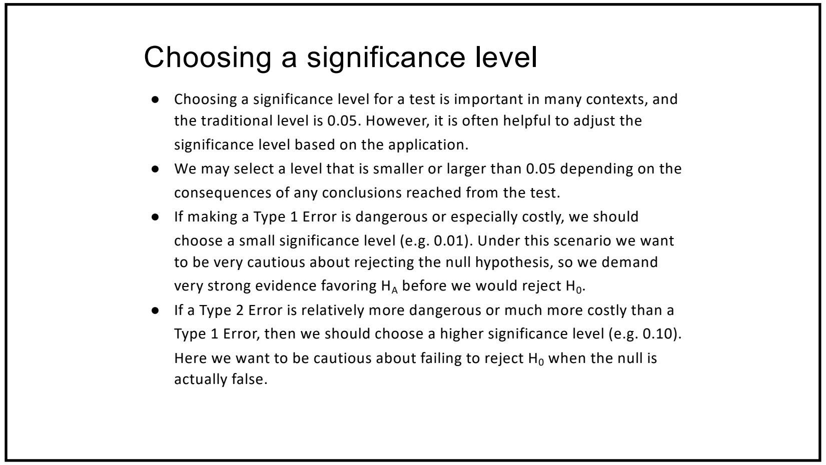
- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error}) = \alpha$$

Or $$P(\text{Reject H0} \mid \text{H0 true}) = \alpha$$

This is why we prefer small values of α -- increasing α increases the Type 1 error rate.

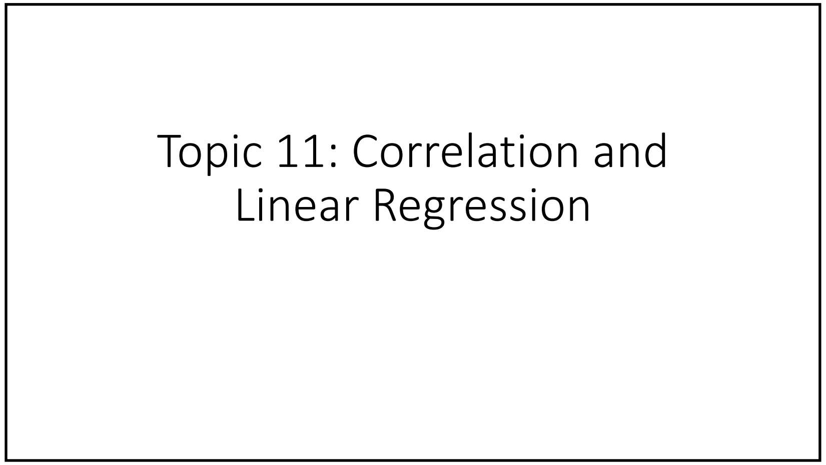
60



Choosing a significance level

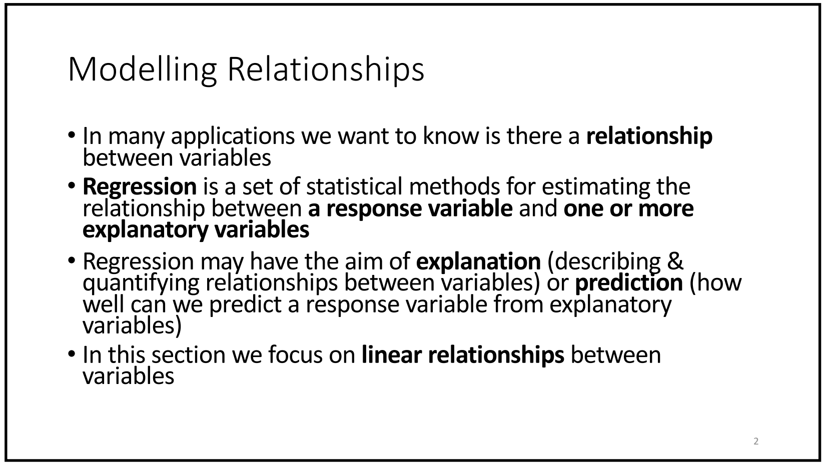
- Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application.
- We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.
- If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.
- If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

61



Topic 11: Correlation and Linear Regression

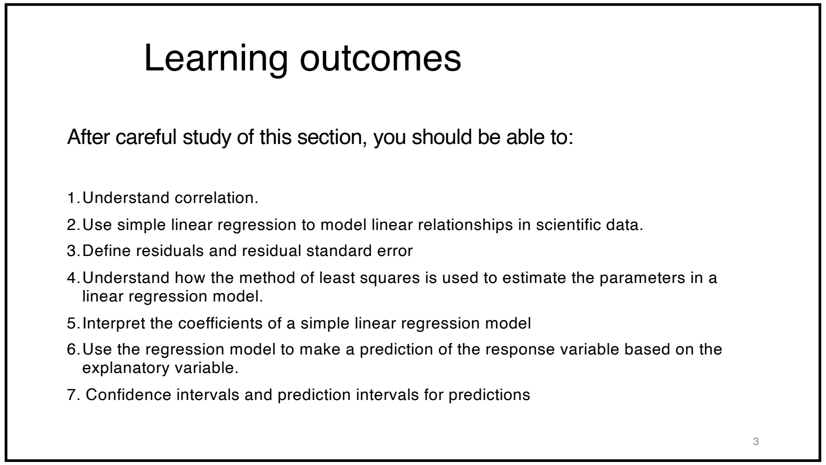
1



Modelling Relationships

- In many applications we want to know is there a **relationship** between variables
- **Regression** is a set of statistical methods for estimating the relationship between **a response variable** and **one or more explanatory variables**
- Regression may have the aim of **explanation** (describing & quantifying relationships between variables) or **prediction** (how well can we predict a response variable from explanatory variables)
- In this section we focus on **linear relationships** between variables

2

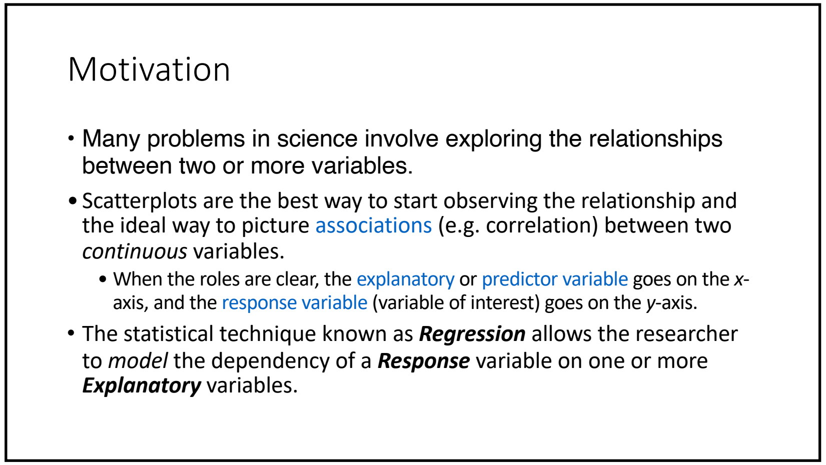


Learning outcomes

After careful study of this section, you should be able to:

1. Understand correlation.
2. Use simple linear regression to model linear relationships in scientific data.
3. Define residuals and residual standard error
4. Understand how the method of least squares is used to estimate the parameters in a linear regression model.
5. Interpret the coefficients of a simple linear regression model
6. Use the regression model to make a prediction of the response variable based on the explanatory variable.
7. Confidence intervals and prediction intervals for predictions

3



Motivation

- Many problems in science involve exploring the relationships between two or more variables.
- Scatterplots are the best way to start observing the relationship and the ideal way to picture associations (e.g. correlation) between two *continuous* variables.
  - When the roles are clear, the explanatory or predictor variable goes on the *x*-axis, and the response variable (variable of interest) goes on the *y*-axis.
- The statistical technique known as ***Regression*** allows the researcher to *model* the dependency of a ***Response*** variable on one or more ***Explanatory*** variables.

4

Motivating Example

- Windfarms are used to generate direct current. Data are collected on 34 different days to determine the relationship between wind speed in mi/h and current in kA.

5

Data:

Name of data file: Windspeed.csv

Response Variable: current in kA

Explanatory Variable: wind speed in mi/h

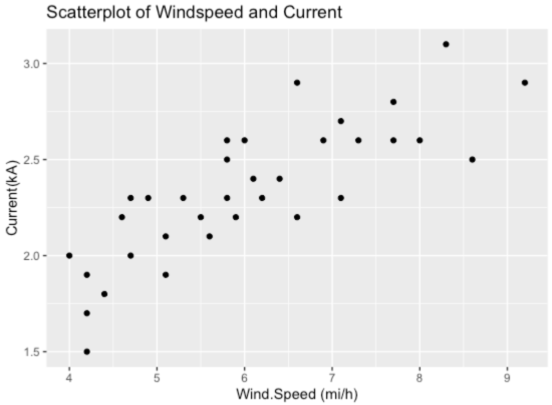
6

```
windspeed.df %>%
  select(Current, Wind.Speed) %>%
  summary()

##     Current          Wind.Speed
##  Min.   :1.500   Min.   :4.000
##  1st Qu.:2.125   1st Qu.:4.950
##  Median :2.300   Median :5.850
##  Mean   :2.335   Mean   :6.047
##  3rd Qu.:2.600   3rd Qu.:7.050
##  Max.   :3.100   Max.   :9.200
```

7

```
ggplot(windspeed.df, aes(y = Current, x = Wind.Speed)) +
  geom_point() +
  labs(x = "Wind.Speed (mi/h)", y = "Current(kA)",
       title = "Scatterplot of Windspeed and Current")
```

8



9

Subjective Impressions ?

- Does it look like there is a relationship between windspeed and current ?
- What is the direction of relationship ?
- How would you quantify the strength of the relationship ?

10

Sample Correlation Coefficient

The sample correlation coefficient *(r)* gives a numerical measurement of the strength of the linear relationship between the explanatory and response variables.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

11

Correlation Coefficient

ρ = +1 means a perfect, linear direct relationship between X and Y

ρ = 0 means no linear relationship between X and Y

ρ = -1 means a perfect, inverse linear relationship between X and Y.

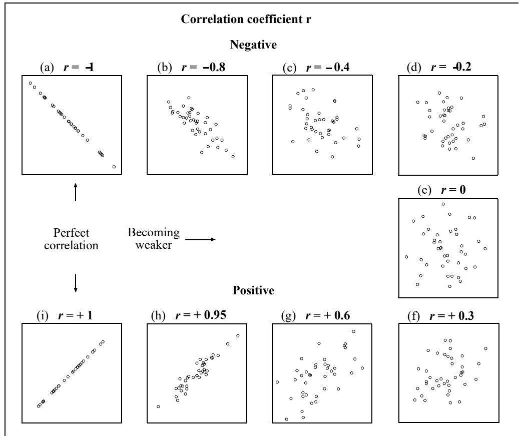
Note: ρ is the population correlation coefficient while r is the sample correlation coefficient.

12

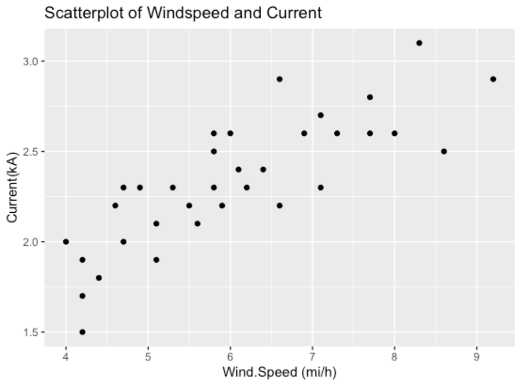
Correlation Coefficient

- Correlation treats *x* and *y* symmetrically:
  - The correlation of *x* with *y* is the same as the correlation of *y* with *x*.
- Correlation has no units.
- Correlation is not affected by changes in the center or scale of either variable.

13



14

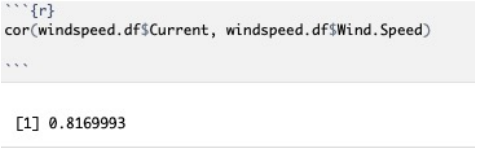


15

Correlation coefficient

```
windspeed.df %>%
  select(Current, Wind.Speed) %>%
  cor()

##              Current Wind.Speed
## Current    1.0000000  0.8169993
## Wind.Speed 0.8169993  1.0000000
```

Or directly using **cor** function as below:



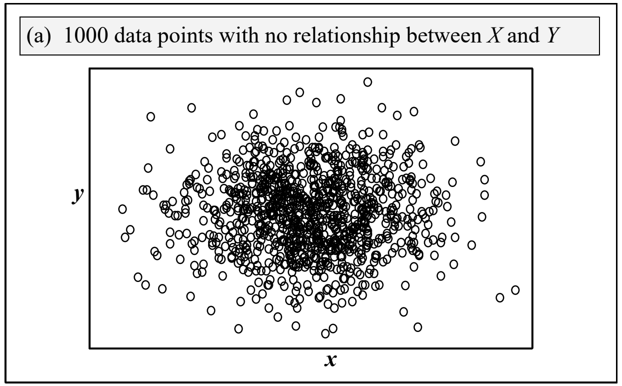
~~~
```{r}
cor(windspeed.df$Current, windspeed.df$Wind.Speed)
```

[1] 0.8169993
~~~

16

Correlation of zero ?

- Sketch what it looks like …

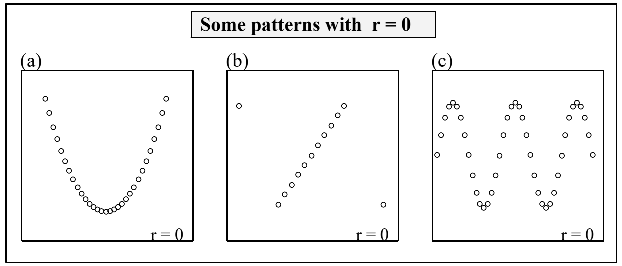
17



(a) 1000 data points with no relationship between X and Y

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 1999.

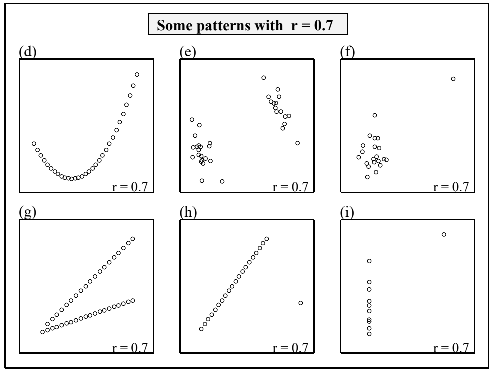
18



Some patterns with r = 0

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

19

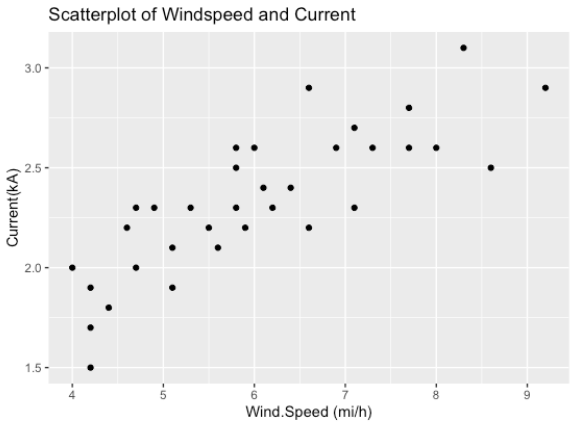


Some patterns with r = 0.7

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

20

r=0.817



21

Take home message …

- Show me the data
- The correlation coefficient measures only linear association
- The correlation coefficient can be misleading when outliers are present
- **Correlation does not imply causation**

22

Correlation ≠ Causation

- Whenever we have a strong correlation, it is tempting to explain it by imagining that the predictor variable has caused the response to help.
- Scatterplots and correlation coefficients never prove causation.
- A hidden variable that stands behind a relationship and determines it by simultaneously affecting the other two variables is called a lurking or confounding variable.

23

Correlation ≠ Causation

- Don’t say “correlation” when you mean “association.
- More often than not, people say correlation when they mean association.
- The word “correlation” should be reserved for measuring the strength and direction of the linear relationship between two quantitative variables.

24

Summary so far ….

- Scatterplots are useful graphical tools for assessing *direction*, *form*, *strength*, and *unusual features* between two variables.
- Although not every relationship is linear, when the scatterplot is straight enough, the *correlation coefficient* is a useful numerical summary.
  - The sign of the correlation tells us the direction of the association.
  - The magnitude of the correlation tells us the *strength* of a linear association.
  - Correlation has no units, so shifting or scaling the data, standardizing, or swapping the variables has no effect on the numerical value.

25

Simple Linear Regression

- Simple linear regression is the name given to the statistical technique that is used to model the dependency of a response variable on a single explanatory variable
  - the word 'simple' refers to the fact that a single explanatory variable is available.
- Simple linear regression is appropriate if the average value of the response variable is a ***linear*** function of the explanatory i.e. the underlying dependency of the response on the explanatory appears linear.

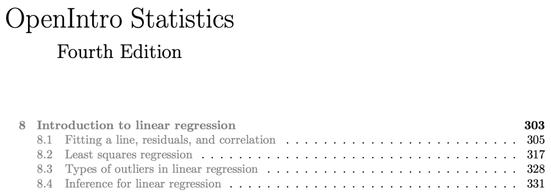
26

Strategy

- Propose a model
- Check the assumptions
- Make some predictions
- Assess how useful it is
- Improve it.

27

Simple Linear Regression

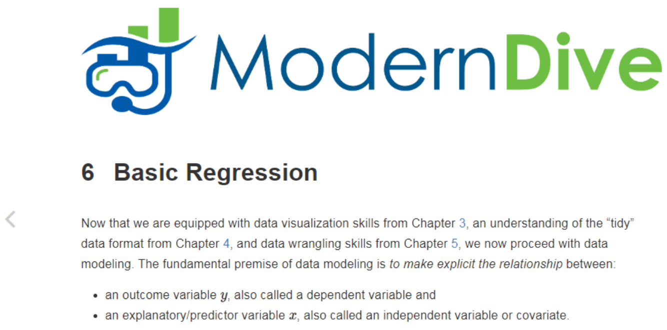


OpenIntro Statistics

Fourth Edition

28

28



ModernDive

6 Basic Regression

Now that we are equipped with data visualization skills from Chapter 3, an understanding of the "tidy" data format from Chapter 4, and data wrangling skills from Chapter 5, we now proceed with data modeling. The fundamental premise of data modeling is *to make explicit the relationship* between:

- an outcome variable y, also called a dependent variable and
- an explanatory/predictor variable x, also called an independent variable or covariate.

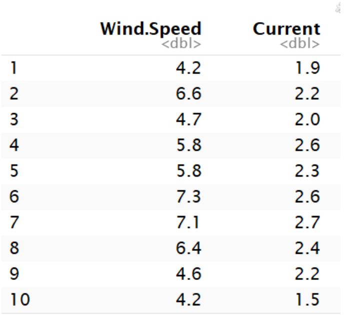
29

Motivating Example

- Windfarms are used to generate direct current. Data are collected on 34 different days to determine the relationship between wind speed in mi/h and current in kA.

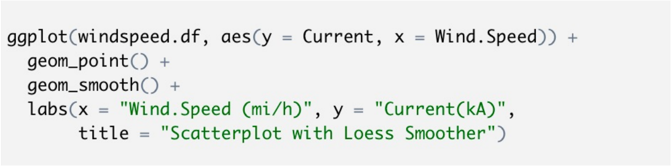
30

A glimpse of the first few rows of data ..



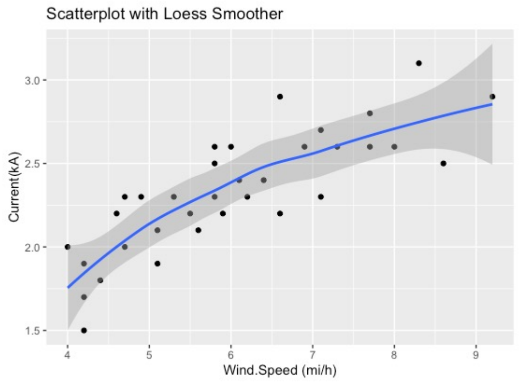
	Wind.Speed <dbl>	Current <dbl>
1	4.2	1.9
2	6.6	2.2
3	4.7	2.0
4	5.8	2.6
5	5.8	2.3
6	7.3	2.6
7	7.1	2.7
8	6.4	2.4
9	4.6	2.2
10	4.2	1.5

31



```
ggplot(windspeed.df, aes(y = Current, x = Wind.Speed)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Wind.Speed (mi/h)", y = "Current(kA)",
       title = "Scatterplot with Loess Smoother")
```

32



33

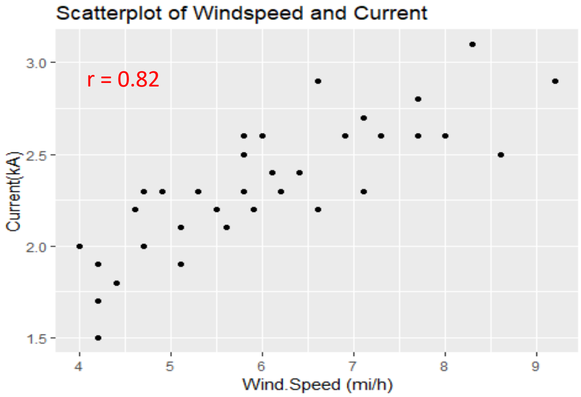
Simple Linear Regression

- The simple linear regression model is of the form

Response Variable = Intercept + Slope*Explanatory Variable
+ random variability

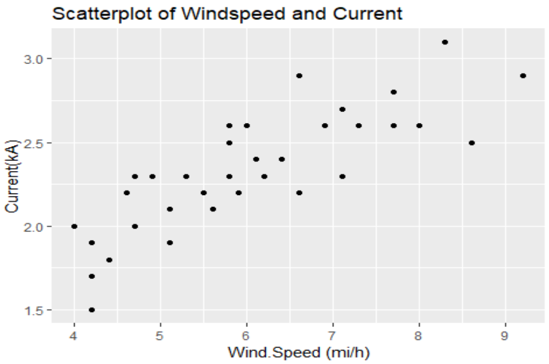
where the intercept and slope must be estimated from a relevant sample of data from the population of interest.

34

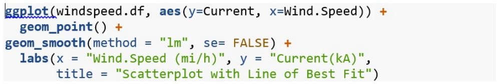


35

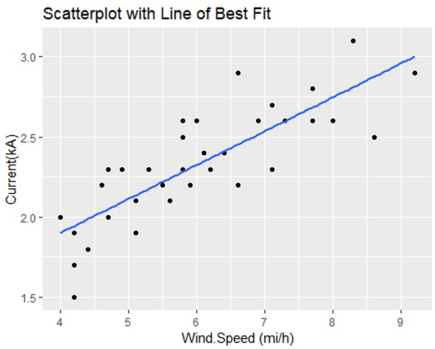
Line of best fit ?



36



```
ggplot(windspeed.df, aes(y=Current, x=Wind.Speed)) +
  geom_point() +
geom_smooth(method = "lm", se= FALSE) +
  labs(x = "Wind.Speed (mi/h)", y = "Current(kA)",
       title = "Scatterplot with Line of Best Fit")
```



37

Interpreting the Slope and Intercept

```
Regression Equation
Mean Current = 1.057 + 0.2113 Wind Speed
```

- b_1 is the slope, which tells us how rapidly $\hat{y}$ changes with respect to x e.g. what is the change in the mean current per unit increase in wind speed.

- b_0 is the y-intercept, which tells where the line crosses (intercepts) the y-axis when x is zero e.g. what is the mean current when wind speed is zero.

38

Predict the Current when Wind Speed = 7.1

```
Regression Equation
Mean Current = 1.057 + 0.2113 Wind Speed
```

39

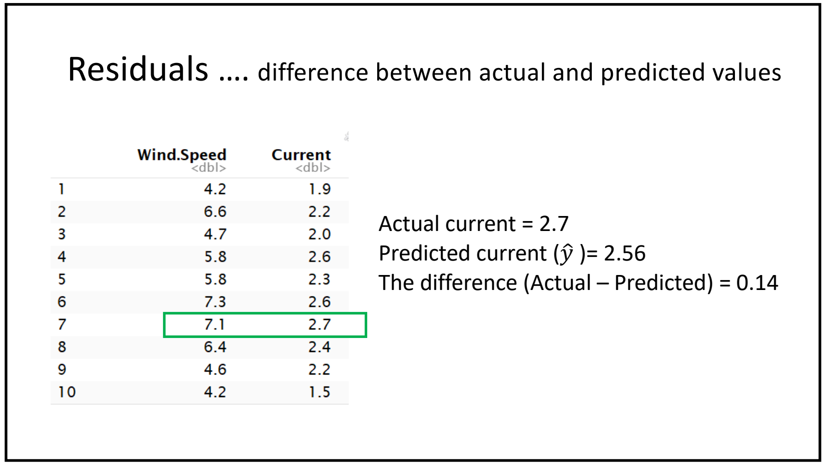
Predict the Current when Wind Speed = 7.1

```
Regression Equation
Mean Current = 1.057 + 0.2113 (7.1) =
2.56
```

The predicted value is often referred to as $\hat{y}$ (i.e. ‘y hat’).

From looking at the data the 7th observation was for a wind speed of 7.1 where the *actual* Current (i.e. y) was equal to 2.7.

40

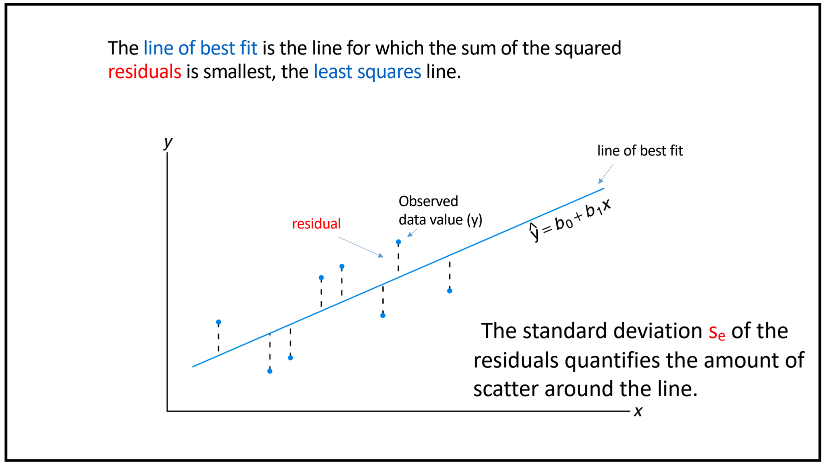


Residuals …. difference between actual and predicted values

	Wind.Speed <dbl>	Current <dbl>
1	4.2	1.9
2	6.6	2.2
3	4.7	2.0
4	5.8	2.6
5	5.8	2.3
6	7.3	2.6
7	7.1	2.7
8	6.4	2.4
9	4.6	2.2
10	4.2	1.5

Actual current = 2.7
Predicted current ($\hat{y}$)= 2.56
The difference (Actual – Predicted) = 0.14

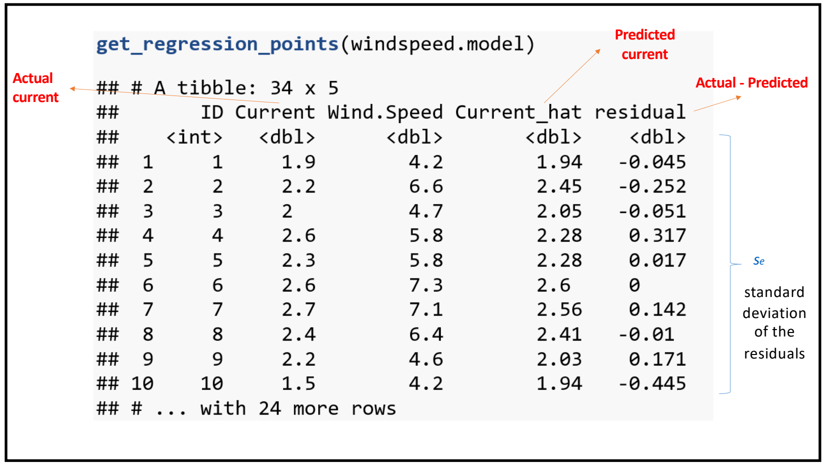
41



The line of best fit is the line for which the sum of the squared residuals is smallest, the least squares line.

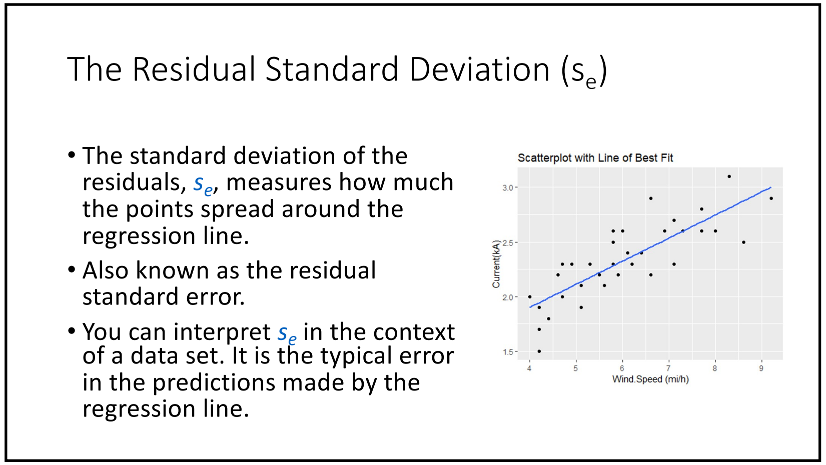
The standard deviation s_e of the residuals quantifies the amount of scatter around the line.

42



```
get_regression_points(windspeed.model)

## # A tibble: 34 x 5
##       ID Current Wind.Speed Current_hat residual
##    <int>   <dbl>      <dbl>       <dbl>    <dbl>
##  1     1     1.9        4.2        1.94   -0.045
##  2     2     2.2        6.6        2.45   -0.252
##  3     3     2          4.7        2.05   -0.051
##  4     4     2.6        5.8        2.28    0.317
##  5     5     2.3        5.8        2.28    0.017
##  6     6     2.6        7.3        2.6     0
##  7     7     2.7        7.1        2.56    0.142
##  8     8     2.4        6.4        2.41   -0.01
##  9     9     2.2        4.6        2.03    0.171
## 10    10     1.5        4.2        1.94   -0.445
## # ... with 24 more rows
```

43



The Residual Standard Deviation (s_e)

- The standard deviation of the residuals, s_e, measures how much the points spread around the regression line.
- Also known as the residual standard error.
- You can interpret s_e in the context of a data set. It is the typical error in the predictions made by the regression line.

44

Line of 'best fit'.

- The line of best fit is the line for which the sum of the squared residuals is smallest, the least squares line.
- Some residuals are positive, others are negative, and, on average, they cancel each other out.
- You can't assess how well the line fits by adding up all the residuals.

45

- Simple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i=1, \ldots, n \text{ assuming } \varepsilon_i \sim N(0, \sigma_e)$$

- Features of this model:
- β_0 (intercept) and β_1 (slope) are the population parameters of the model and must be estimated from the data as b_0 (sample intercept) and b_1 (sample slope).
- The process of estimating β_0 and β_1 is called fitting the model to the data.
- $\beta_0 + \beta_1 x_i$ is the population mean response (mean of *Y*) given $X=x_i$.
- ε_i is the error term in the regression model. Actually it refers to the difference between the fitted line and y_i.
- σ_e (error) is the stochastic part of the model (unexplained variability). Or in other words, it is the standard deviation corresponding to the error term.
- Once estimated predicted values for y (labelled as $\hat{y}$) can be made as follows:

$$\hat{y} = b_0 + b_1 x$$

- $\hat{y}$ is used to emphasize that the points that satisfy this equation are just our *predicted* values, not the actual data values.

46

Estimating the Slope (least squares)

- In the simple linear regression model the slope (b_1) is built from the correlation coefficient r and the standard deviations of y and x:

$$b_1 = r\frac{s_y}{s_x}$$

- The slope is always in units of *y* per unit of *x*.

47

Estimating the Intercept (least squares)

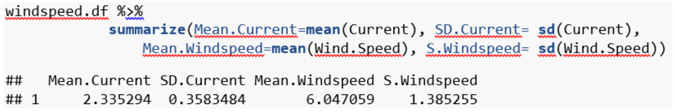
- In the simple linear regression model the intercept (b_0) the intercept is built from the means and the slope:

$$b_0 = \bar{y} - b_1\bar{x}$$

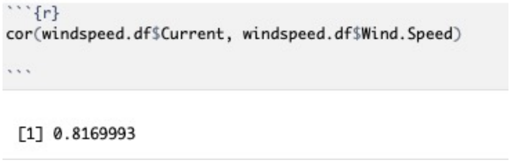
- The intercept is always in units of *y*.
- We almost always use technology to find the equation of the regression line.

48

Summary Statistics



```
windspeed.df %>%
            summarize(Mean.Current=mean(Current), SD.Current= sd(Current),
                Mean.Windspeed=mean(Wind.Speed), S.Windspeed= sd(Wind.Speed))

##    Mean.Current SD.Current Mean.Windspeed S.Windspeed
## 1      2.335294  0.3583484       6.047059    1.385255
```



~~~
```{r}
cor(windspeed.df$Current, windspeed.df$Wind.Speed)

```

 [1] 0.8169993
~~~

49

Slope and Intercept

$b_1 = r\frac{s_y}{s_x}$ = ,

$b_0 = y - b_1 x =$

Regression Equation
Mean Current = + Wind Speed

50

Slope and Intercept

$b_1 = r\frac{s_y}{s_x}$ = 0.817$\frac{0.3583}{1.385}$ = 0.2113,

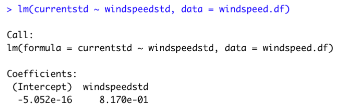
$b_0 = y - b_1 x =$ 2.3353 −0.2113(6.047) = 1.057

Regression Equation
Mean Current = 1.057 + 0.2113 Wind Speed

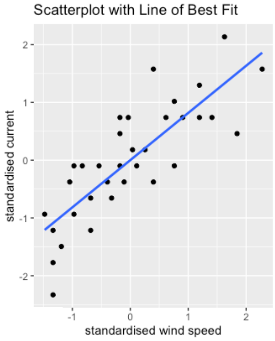
51

Windspeed and current standardized (subtract mean and divide by sd)
So standardized values have mean 0 and sd 1

r = 0.817



```
> lm(currentstd ~ windspeedstd, data = windspeed.df)

Call:
lm(formula = currentstd ~ windspeedstd, data = windspeed.df)

Coefficients:
 (Intercept)  windspeedstd
  -5.052e-16     8.170e-01
```



Scatterplot with Line of Best Fit

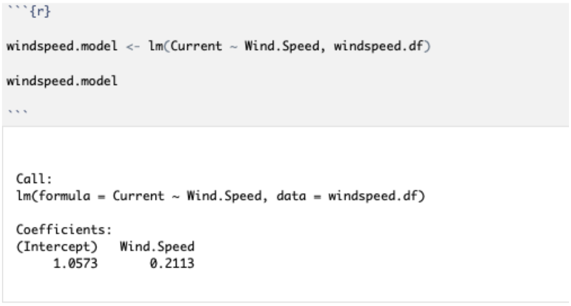
52

Summary so far …

- **Correlation** is a useful metric for measuring the degree of **linear relationship** between two continuous variables
- **Regression** is a useful tool for **modelling** the relationship between two continuous variables: a response (y) and an explanatory/predictor (x)
- The line of best fit is the line where the sum of the squared residuals (difference between observed and fitted values) is a minimum
- To use this line to make **inference** (and predictions) there are several **assumptions** that must be satisfied

53

Fitting a Simple Linear Regression in R

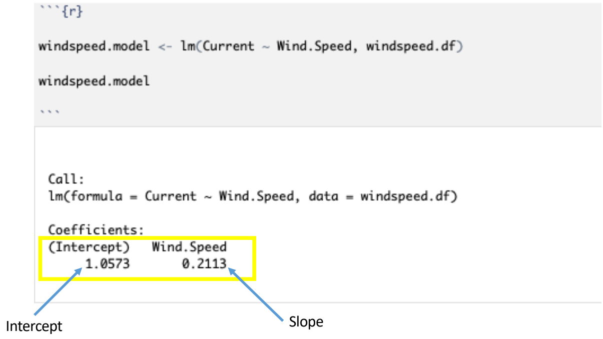


~~~
```{r}
windspeed.model <- lm(Current ~ Wind.Speed, windspeed.df)
windspeed.model
```
~~~

```
Call:
lm(formula = Current ~ Wind.Speed, data = windspeed.df)

Coefficients:
(Intercept)   Wind.Speed  
     1.0573       0.2113  
```

54

Fitting a Simple Linear Regression in R



~~~
```{r}
windspeed.model <- lm(Current ~ Wind.Speed, windspeed.df)
windspeed.model
```
~~~

```
Call:
lm(formula = Current ~ Wind.Speed, data = windspeed.df)

Coefficients:
(Intercept)   Wind.Speed  
     1.0573       0.2113  
```

55

Interpreting the Slope and Intercept

Regression Equation
Mean Current = 1.057 + 0.2113 Wind Speed

- b_1 is the slope, which tells us how rapidly $\hat{y}$ changes with respect to x e.g. what is the change in the (mean) current per unit increase in wind speed.
- b_0 is the *y*-intercept, which tells where the line crosses (intercepts) the y-axis when x is zero e.g. what is the (mean) current when wind speed is zero.

56

Inference for predictions

- We have seen how to make **point estimates** of the predicted response
- Just as in inference for the true mean, an interval estimate is more useful for inference
- We look at two types of **interval estimates for the mean (or predicted) response given some value of the explanatory variable**
- 1. Confidence interval
- 2. Prediction interval

57

Confidence Interval for the mean response

- A range of values that is likely to contain the **true mean value of the response variable given a specific values of the the explanatory variable**.
- This range **doesn't tell** you about the spread of the **individual data points** around the true mean.

58

Prediction Interval for response in new observations

- A range of values that is likely to contains the value of the response variable for a **single new observation** given a specific value of the explanatory variable.
- The prediction interval is for **individual observations rather than the mean**.

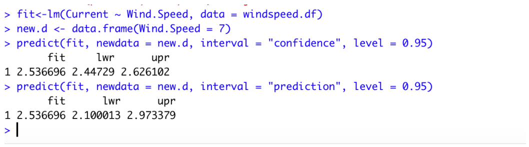
59

For prediction in R: the predict() function

- predict(object, newdata, se.fit = FALSE, interval = c("none", "confidence", "prediction"), level = 0.95)
- object a fitted lm() model object.
- newdata An optional data frame in which to look for variables with which to predict.
- se.fit A switch indicating if standard errors for predictions are required. The default is se.fit = FALSE.
- interval Type of interval to be calculated. The default is interval = "none".
- level the confidence level for generating interval estimates. The default is level = 0.95.

60

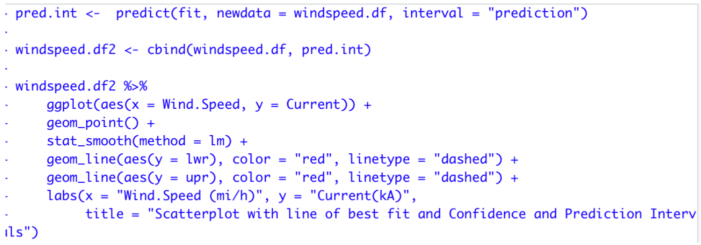
R code for confidence interval and prediction interval for a single point



```
> fit<-lm(Current ~ Wind.Speed, data = windspeed.df)
> new.d <- data.frame(Wind.Speed = 7)
> predict(fit, newdata = new.d, interval = "confidence", level = 0.95)
       fit     lwr      upr
1 2.536696 2.44729 2.626102
> predict(fit, newdata = new.d, interval = "prediction", level = 0.95)
       fit      lwr      upr
1 2.536696 2.100013 2.973379
> |
```

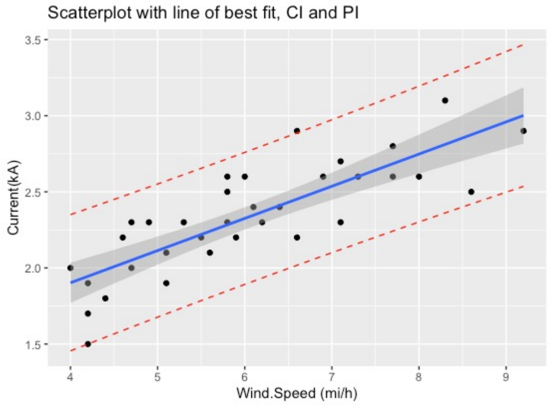
61

R code for pointwise CI and PI



```
· pred.int <-  predict(fit, newdata = windspeed.df, interval = "prediction")
·
· windspeed.df2 <- cbind(windspeed.df, pred.int)
·
· windspeed.df2 %>%
·     ggplot(aes(x = Wind.Speed, y = Current)) +
·     geom_point() +
·     stat_smooth(method = lm) +
·     geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
·     geom_line(aes(y = upr), color = "red", linetype = "dashed") +
·     labs(x = "Wind.Speed (mi/h)", y = "Current(kA)",
·          title = "Scatterplot with line of best fit and Confidence and Prediction Interv
ıls")
```

62

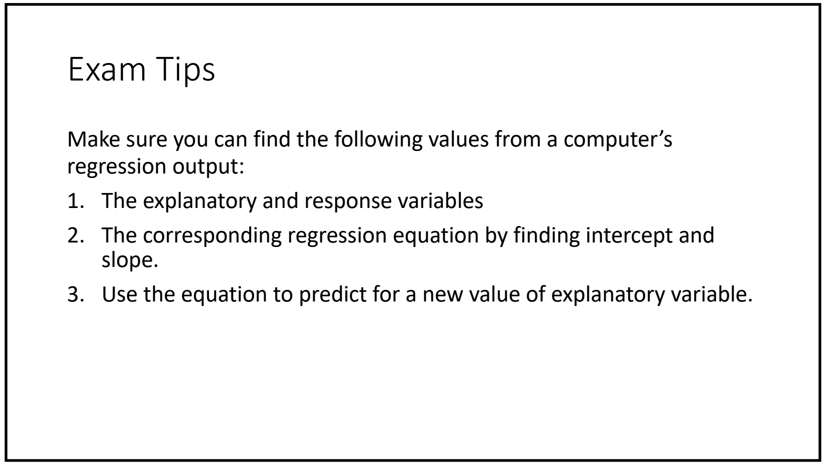


63

What Can Go Wrong?

- Don’t fit a straight line to a nonlinear relationship.
- Beware extraordinary points (**y-values that stand off from the linear pattern or extreme *x*-values**).
- Don’t extrapolate beyond the data—the linear model may no longer hold outside of **the range of the data**.
- Don’t infer that *x* causes *y* just because there is a good linear model for their relationship—association is *not* causation.
- An empirical model is valid only for the data to which it is fit. It may or may not be useful in predicting outcomes for subsequent observations.

64



Exam Tips

Make sure you can find the following values from a computer's regression output:

1. The explanatory and response variables
2. The corresponding regression equation by finding intercept and slope.
3. Use the equation to predict for a new value of explanatory variable.

65