

# Data analysis and the R statistical package

---

## Empirical Methods in Software Engineering (01 OPJIU)

<http://softeng.polito.it/EMSE/>



**SoftEng**  
<http://softeng.polito.it>

Version 1.7  
© Marco Torchiano, 2014






This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-nd/4.0/>.

You are free: to copy, distribute, display, and perform the work

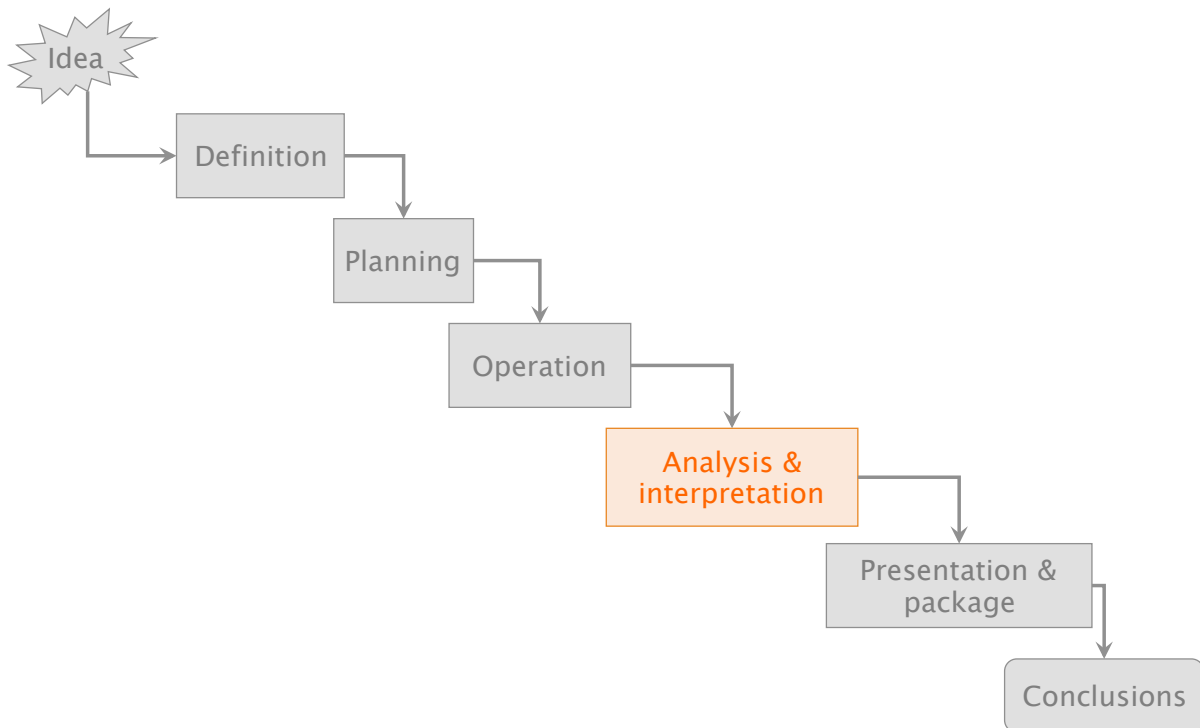
Under the following conditions:

-  **Attribution.** You must attribute the work in the manner specified by the author or licensor.
-  **Non-commercial.** You may not use this work for commercial purposes.
-  **No Derivative Works.** You may not alter, transform, or build upon this work.
  - For any reuse or distribution, you must make clear to others the license terms of this work.
  - Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

# Experimental process

---



## Agenda

---

- R statistical package
- Distributions
  - ◆ Functions
  - ◆ Central Limit Theorem
- Hypothesis testing
  - ◆ One sample
  - ◆ Two samples
  - ◆ Nonparametric tests
  - ◆ ANOVA

---

# R STATISTICAL PACKAGE

## What is R?

---



<http://cran.r-project.org/>

- R is a free software environment for statistical computing and graphics.
  - ♦ Available on several different platform

# CLI

---

- Command Line Interface
  - ♦ Immediate evaluation of expression
- Scripts
  
- Extensive help system
  - ♦ <http://www.rseek.org/>

# GUI

---

- Several GUI front-ends
- RStudio is a full IDE for R
  - ♦ <http://www.rstudio.com>



# R elements

---

- Functions
- Data types
  - ◆ Primitive
  - ◆ Compound

# Functions

---

- Definition
  - ◆ `percentage <-  
function(part,whole) {  
 part / whole * 100;  
}`
- Usual invocation (positional)
  - ◆ `percentage(3, 4)`
- Named arguments
  - ◆ `percentage(whole=4, part=3)`

# Data types

---

- Primitive
- Compound
- Type functions:
  - ◆ Type of variable: `class(x)`
  - ◆ Check type: `is.type(x)`
  - ◆ Conversion: `as.type(x)`

# Primitive types

---

- Numeric
  - ◆ Default type
- Integer
- Complex
  - ◆ `sqrt( 1 + 0i ) → 0 + 1i`
- Logical
  - ◆ `TRUE | FALSE`
- Character
  - ◆ Strings

# Compound types

---

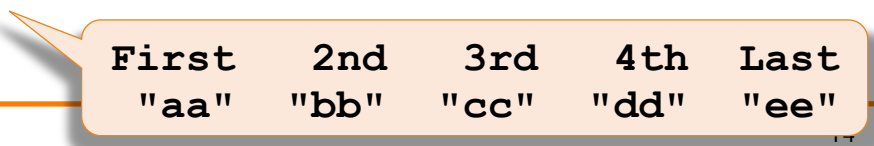
- Vector
  - ♦ Creation: `v = c(2, 4, 5)`
    - Everything is vector: `1 == c(1)`
  - ♦ Sequences: `v = seq(from=1, to=2, by=0.1)`  
`v = rep(1, time=10)`
  - ♦ Range: `1:3` equals to `c(1, 2, 3)`
  - ♦ Merging: `m = c(v1, v2)`
    - Type coercion can be applied
  - ♦ Arithmetic
    - Pair-wise on same-index elements
    - Recycling if different size
      - Longest length must be multiple of smallest length

# Compound types

---

```
s = c("aa", "bb", "cc", "dd", "ee")
```

- Vector index
  - ♦ Operator `[]`
    - Simple index: `s[3] → "cc"`
    - **Slicing** vector index: `s[ c(1, 3) ] → "aa" "cc"`
    - `s[c(5, 1, 1, 3)] → "ee" "aa" "aa" "cc"`
    - Logical vector: `s[ 1 ] → "aa" "ee"`
      - `1 = c(TRUE, FALSE, FALSE, FALSE, TRUE)`
  - ♦ Named vectors
    - `names(s) = c("First", "2nd", "3rd", "4th", "Last")`








First	2nd	3rd	4th	Last
"aa"	"bb"	"cc"	"dd"	"ee"




# Compound types

---

## ▪ Matrix

- ◆ Construction:  $A = \text{matrix}(1:9, 3, 3)$    
- `byrow = FALSE` by default 
- ◆ Transposition:  $B = \text{t}(A)$  
- ◆ Composition:  $C = \text{cbind}(A, B)$    
 $D = \text{rbind}(A, B)$  

## ▪ Index

- ◆ Single element:  $A[1, 3]$  
- ◆ Row:  $A[1, ]$  
- ◆ Column:  $A[, 2]$  

# Compound types

---

## ▪ List

- ◆ An array whose elements can be either primitive or compound types
- ◆ Construction:  $l = \text{list}(c(1,2), "a")$
- ◆ Slicing:  $l[2] \rightarrow \begin{bmatrix} [1] \\ [1] "a" \end{bmatrix}$
- ◆ Member:  $l[[1]] \rightarrow [1] \ 1 \ 2$
- ◆ Named members:
  - Definition:  $l = \text{list}(n=c(1,2), \text{char}="a")$
  - Access:  $l\$n == l[["n"]] == l[[1]]$



# Compound types

---

- Factor
  - ♦ Represent nominal and ordinal variables
    - Internally stored as integer vector
  - ♦ `x = c("A", "B", "B", "D", "A", "D")`
  - ♦ Create: `f = factor(x, levels=c("A", "B", "C", "D"), ordered=T)`
  - ♦ Levels: `levels(f)`
  - ♦ Frequencies: `table(f)`

## Compound

			mpg	cyl	disp	hp
Mazda	RX4		21.0	6	160	110
Mazda	RX4	Wag	21.0	6	160	110

- Dataframe
  - ♦ List of vectors of equal length
  - ♦ Construction: `df = data.frame(...)`
  - ♦ Cell indexing: `df[1,2]`
  - ♦ Column selection:
    - `df[[1]] == df[["mpg"]] == df$mpg`
  - ♦ Dataframe slicing
    - Column: `df[1] == df["mpg"]`
    - Row: `df[1, ] == df["Mazda RX4", ]`  
`== df[, c(TRUE, FALSE)]`

# Import

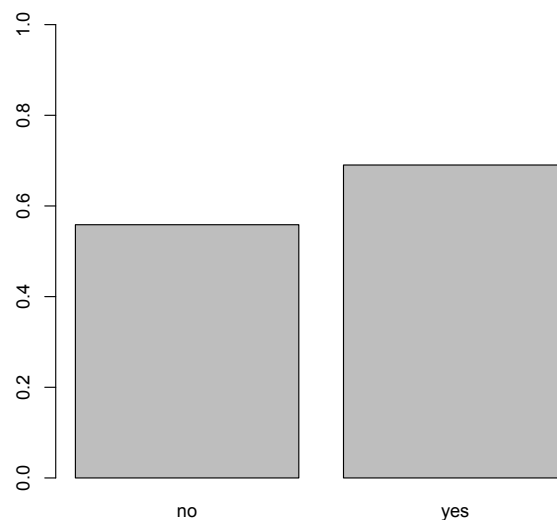
---

- **read.\***
  - ◆ Read data from a file into a dataframe
  - ◆ Space separated: **read.table()**
  - ◆ CSV: **read.csv()**
  - ◆ Clipboard: **read.table(pipe(...))**
    - X11: "clipboard"
    - OS X: "pbpaste"
  - ◆ Excel file
    - **library(gdata)**
    - **read.xls()**

# Diagrams

---

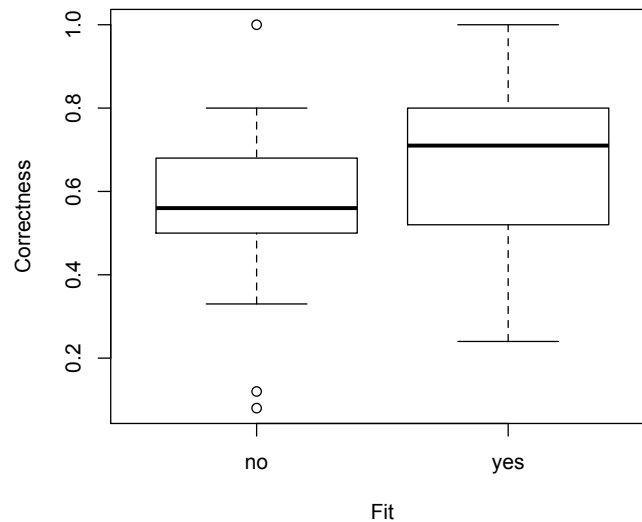
```
barplot(height = s$Complete,  
names.arg=s$Fit, ylim=c(0,1))
```



# Diagrams

---

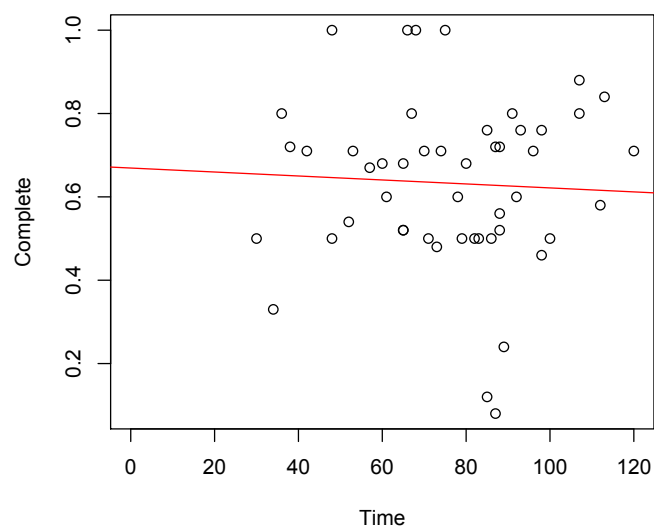
```
boxplot(Complete ~ Fit, data=data,  
        xlab="Fit", ylab="Correctness")
```



# Diagrams

---

```
plot(data$Complete ~ data$TimeTotal,  
      xlim=c(0,120), xlab="Time", ylab="Complete")  
abline(m$coefficients, col="red")
```



---

# DATA FRAMES

## Data frame

---

Variables

Observations

Subject	Treatment	Score
1	Control	9
2	Experiment	8
3	Control	10
4	Experiment	5

# Variables and Categories

---

- Outcome of experiment or questionnaire consist of several observations
- Each observation is characterized by
  - ♦ Quantitative values → the measure of interest
  - ♦ Categorical values →
    - Factors, groups, and blocks
    - Name of variables

## Example

---

- G.Scanniello, F.Ricca, M.Torchiano, G.Reggio, E.Astesiano “Assessing the Effect of Screen Mockups on the Comprehension of Functional Requirements” ACM TRANSACTIONS ON SOFTWARE ENGINEERING AND METHODOLOGY, Vol.24, pp. 1:1–1:38, 2014

# Case Study

Table VI. Postexperiment Survey Questionnaire

Item ID	Question	Valid Answers
PQ1	I had enough time to perform the tasks.	(1-5)
PQ2	The questions of the comprehension questionnaire were clear to me.	(1-5)
PQ3	I did not have any issue in comprehending the use cases.	(1-5)
PQ4	I did not have any issue in comprehending the use case diagrams.	(1-5)
PQ5	I found the exercise useful.	(1-5)
PQ6	I found screen mockups useful (when present)	(1-5)
PQ7	To see the screen mockups (when present), I spent (in terms of percentage) with respect to the total time to accomplish the task.	(A-E)

(1) strongly agree, (2) agree, (3) neither agree nor disagree, (4) disagree, (5) strongly disagree.  
 (A) < 20%, (B) > 20% and ≤ 40%, (C) > 40% and ≤ 60%, (D) > 60% and ≤ 80%, (E) ≥ 80%

—The requirements specification documents in electronic format (MS Word) of Easy-

# Wide Format

Exp	Subject	Group	Q1	Q2	Q3	Q4	Q5	Q6	Q7
1	101	1	1	3	1	1	1	2	5
1	104	1	1	3	2	2	1	2	3
1	108	1	1	3	4	4	1	1	2
1	112	1	1	3	2	2	1	1	3
1	116	1	2	3	2	3	1	2	1
1	120	1	1	2	2	2	1	1	4
1	122	1	1	2	1	2	1	1	2

# Wide Format

Exp	Subject	Group	Q1	Q2	Q3	Q4	Q5	Q6	Q7
1	101	1	1	3	1	1	1	2	5
1	104	1	1	3	2	2	1	2	3
1	108	1	1	3	2	2	1	1	2
1	112	1	1	3	2	2	1	1	3
1	116	1	2	3	2	3	1	2	1
1	120	1	1	2	2	2	1	1	4
1	122	1	1	2	1	2	1	1	2

Annotations in the image: 'ID vars' points to the Subject column, and 'Measure vars' points to the Q1-Q7 columns.

# Long Format

Exp	Subject	Group	variable	value
1	101	1	Q1	1
1	104	1	Q1	1
1	108	1	Q1	1
1	112	1	Q1	1
1	116	1	Q1	2
1	120	1	Q1	1
1	122	1	Q1	1

# Conversion

---

```
require(reshape2)
Dw =read.table("MockupPostQuest.csv")
## from wide to long
dl = melt(dw, id.vars=
          .(Exp, Subject, Group))
## from long to wide
dw = dcast(dl, ... ~ variable)
```

# Example

---

- Data from
  - ♦ Are Fit Tables Really Talking?  
A Series of Experiments to Understand whether Fit Tables are Useful during Evolution Tasks
  - ♦ <http://www.rcost.unisannio.it/mdipenta/Fit-Package.zip>



---

# DISTRIBUTIONS

## Distributions

---

- Probability distribution describes the probability of a random variable to assume certain values
  - ♦ Discrete
  - ♦ Continuous

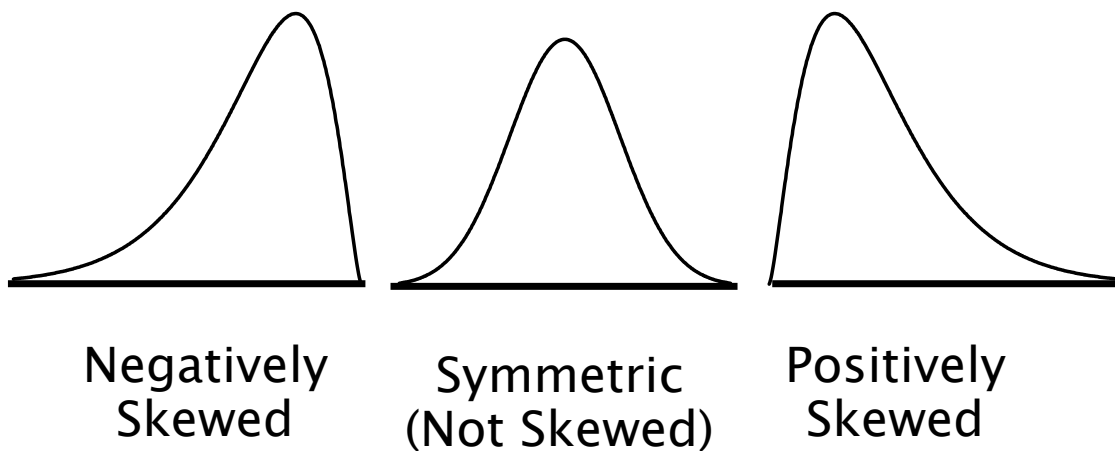
# Measures of Shape

---

- **Skewness**: absence of symmetry
  - ◆ Extreme values in one side of a distribution
- **Kurtosis**: peakedness of a distribution
  - ◆ Leptokurtic: high and thin
  - ◆ Mesokurtic: normal shape
  - ◆ Platykurtic: flat and spread out

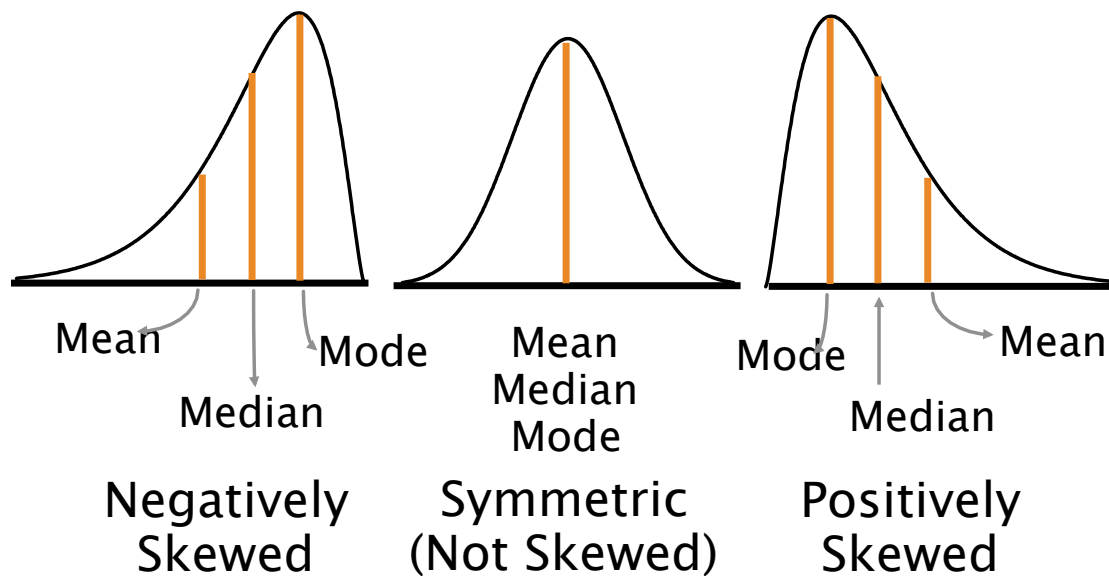
## Skewness

---



# Skewness

---



## Coefficient of Skewness

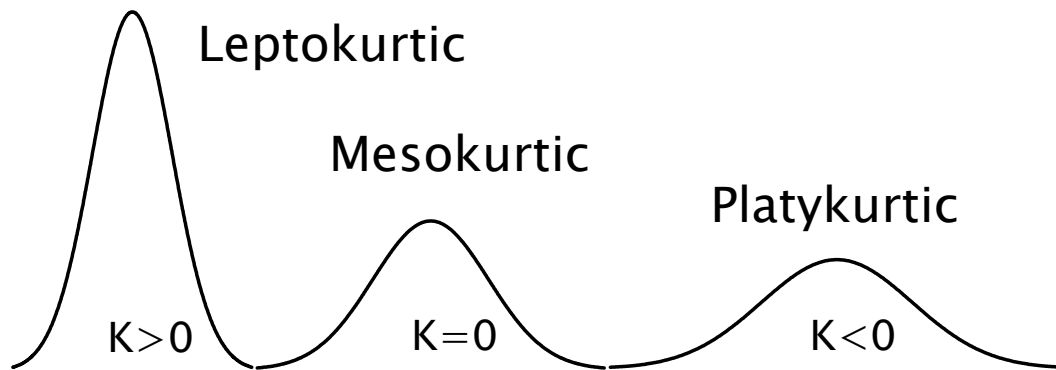
---

- S: Summary measure for skewness
- **library (moments)**  
**S = skewness (x)**
  - ♦ If  $S < 0$ , the distribution is negatively skewed (skewed to the left).
  - ♦ If  $S = 0$ , the distribution is symmetric (not skewed).
  - ♦ If  $S > 0$ , the distribution is positively skewed (skewed to the right).

# Kurtosis

---

- Peakedness of a distribution
  - ♦ Leptokurtic: high and thin
  - ♦ Mesokurtic: normal in shape
  - ♦ Platykurtic: flat and spread out



# Kurtosis

---

- $K$ : measure of kurtosis
  - ♦ **library (moments)**  
**kurtosis (x)**
  - ♦  $K > 0$ : leptokurtic
  - ♦  $K = 0$ : mesokurtic
  - ♦  $K < 0$ : platykurtic

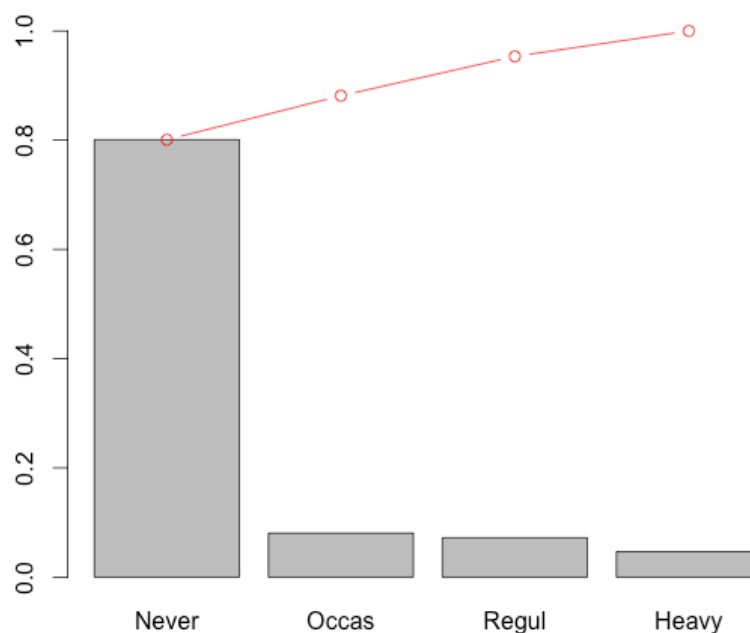
# Discrete distribution

---

- Absolute Frequency
  - ♦ `table(v)`
  - ♦ `barplot(table(v))`
- Relative frequency
  - ♦ `table(v) / length(v)`
- Cumulative frequency
  - ♦ `cumsum(table(v))`
    - Meaningful for at least ordinal data

# Density + Cumulative

---



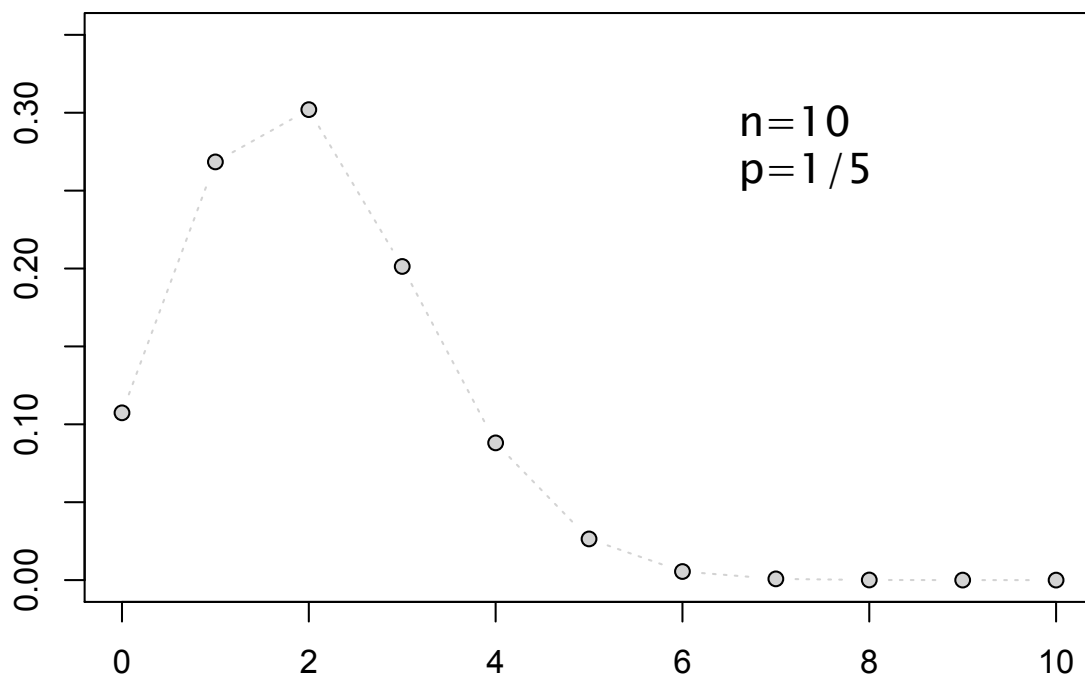
# Binomial

---

- Describes the outcome of  $n$  independent trials in an experiment. Each trial is assumed to have only two outcome, labeled as success or failure
  - ♦ Probability of having  $x$  successful trials

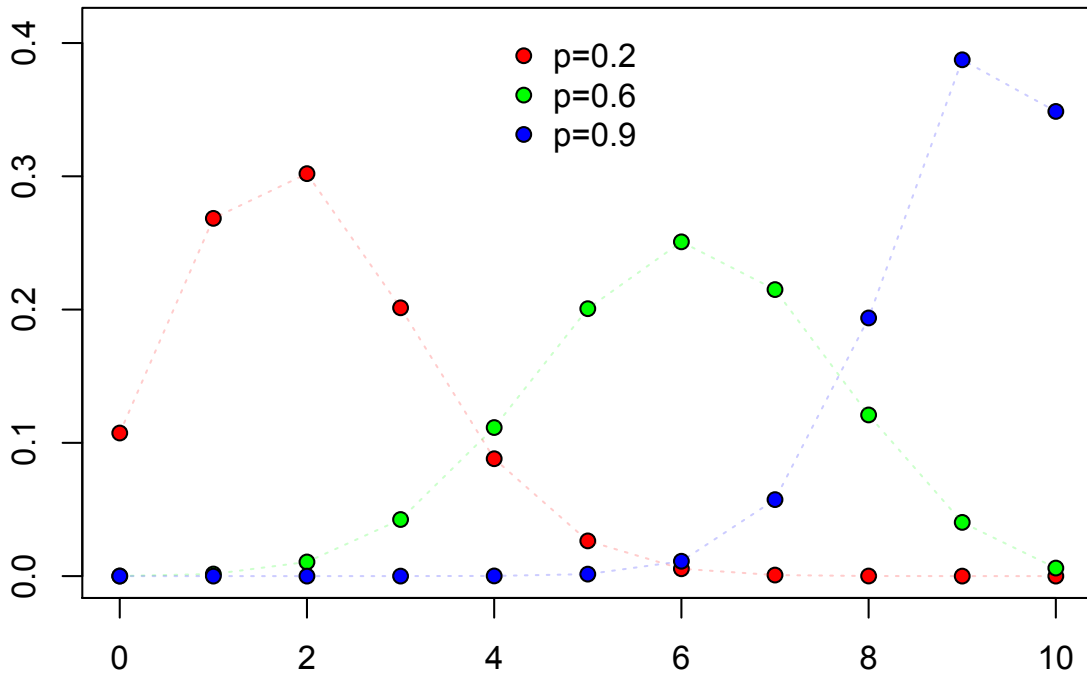
## Binomial – PMF

---



# Binomial – PMF

---

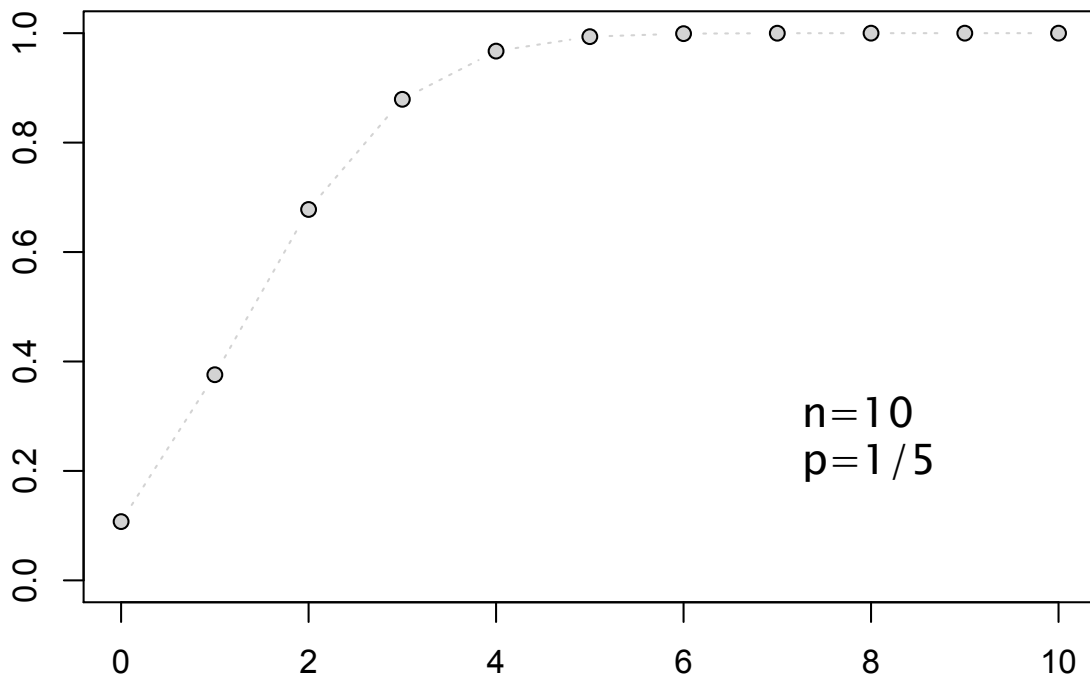


# Binomial – Example

---

- **Ten** multiple choice questions in a quiz. Each question has **five** possible answers, and only **one** of them is correct.
- Find the probability of having **four** or less correct answers if a student attempts to answer every question at random.
  - ♦ `pbinom(4, size=10, prob=1/5)`

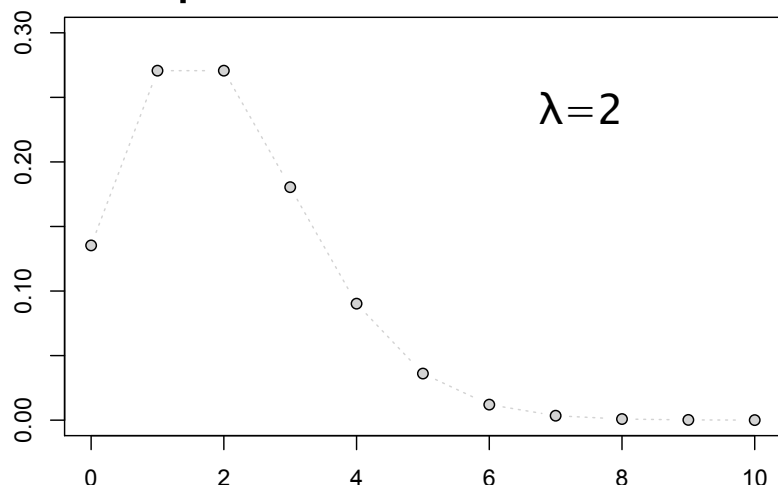
# Binomial – CDF



# Poisson

- Distribution of independent events occurrence in an interval. If  $\lambda$  is the mean occurrence per interval.

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$





# Normal

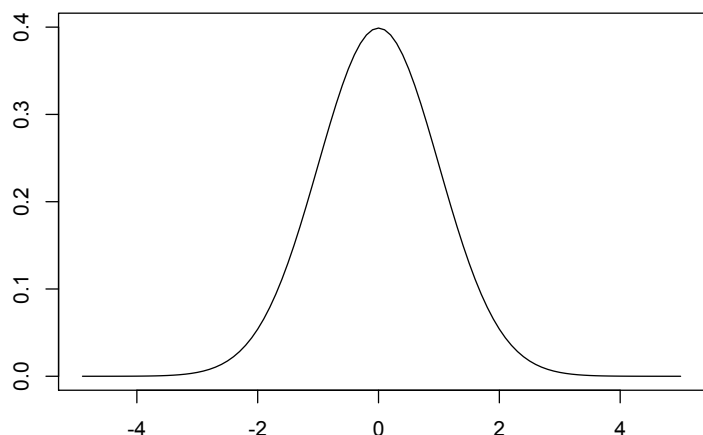
---

- Continuous distribution
- Symmetrical distribution
- Asymptotic to the horizontal axis
- Unimodal
- A family of curves
- Area under the curve sums to 1.

# Normal

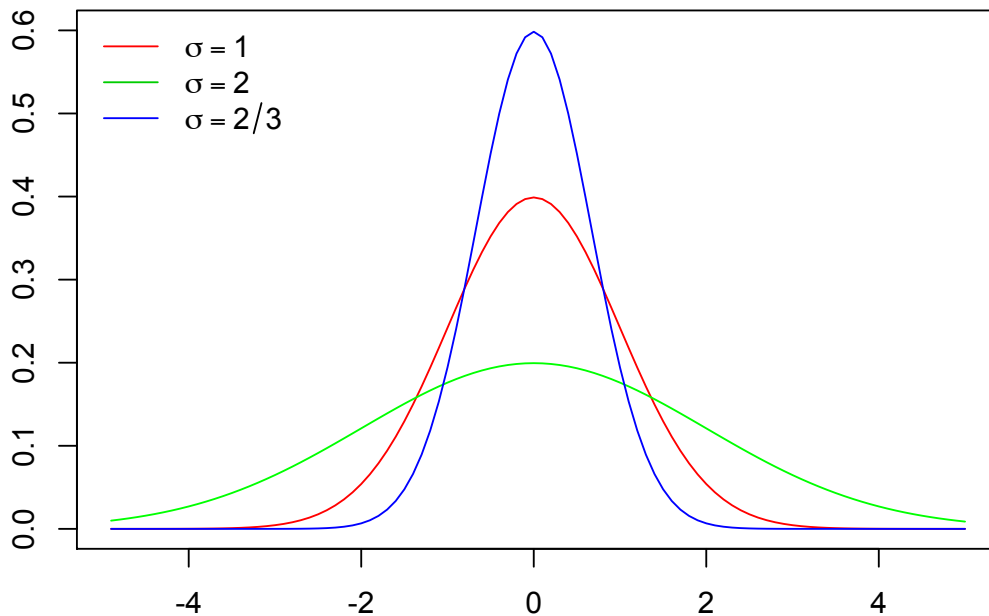
---

- Equation:  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ 
  - ♦  $\mu$  = mean of X
  - ♦  $\sigma$  = standard deviation of X
  - ♦  $\pi = 3.13159\dots$
  - ♦  $e = 2.71828\dots$



# Normal

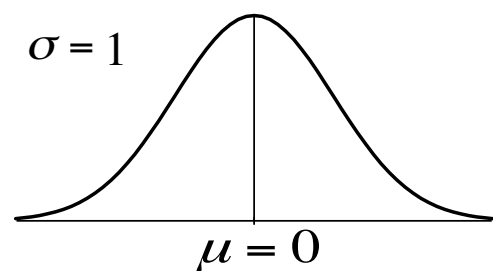
---



# Standardized Normal

---

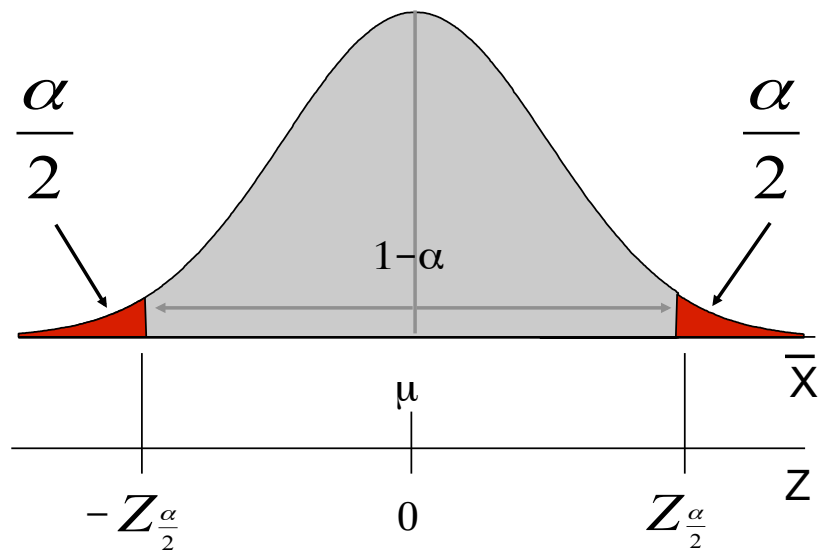
- Distribution with
  - ♦ a mean of zero, and
  - ♦ a standard deviation of one
- Z Formula
  - ♦ standardizes any normal distribution
- Z Score
  - ♦ computed by the Z Formula
  - ♦ the number of standard deviations which a value is away from the mean



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

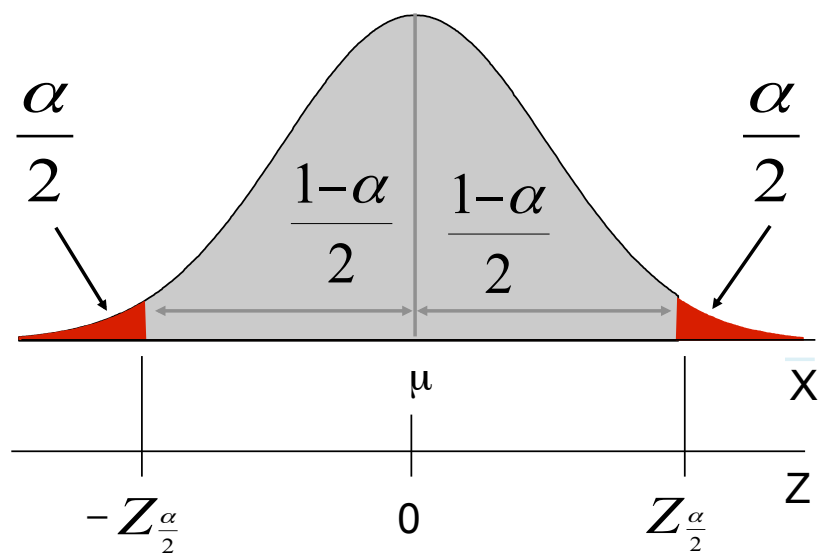
# $1-\alpha$ confidence interval

---

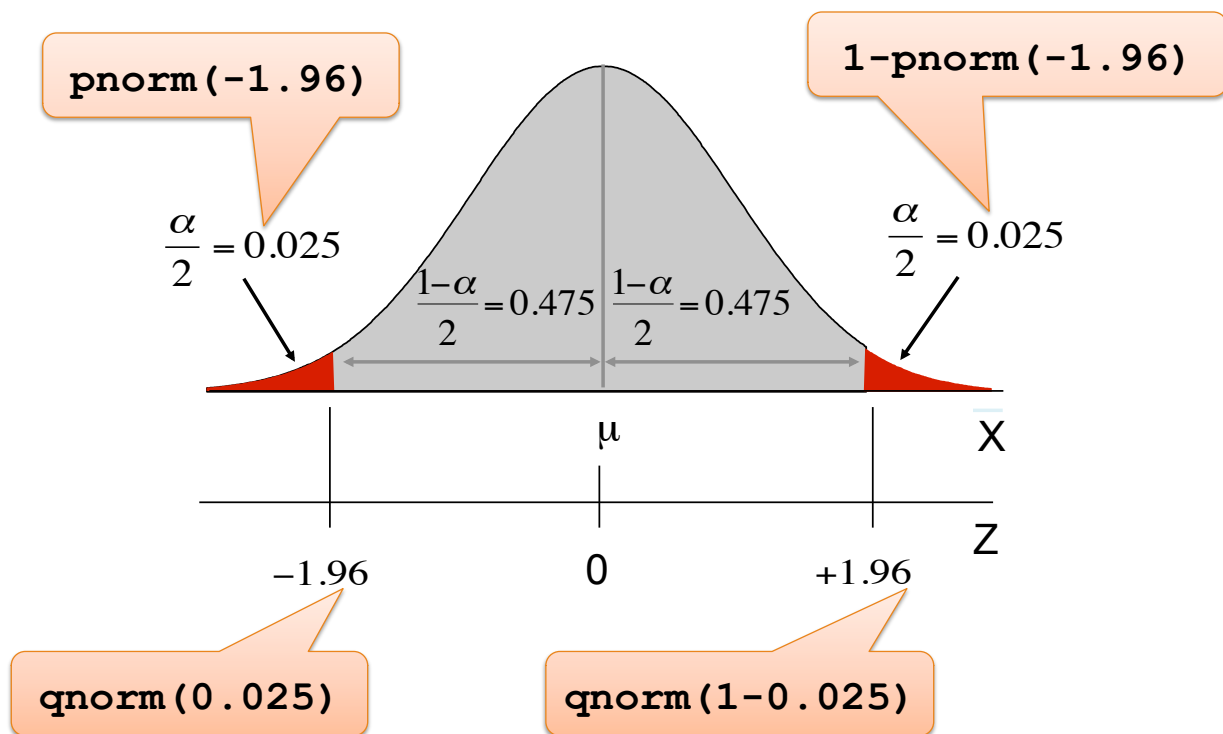


# $1-\alpha$ confidence interval

---



# 95% confidence interval



## Summary of confidence levels

Confidence Level	Z Value
90%	1.645
95%	1.96
98%	2.326
99%	2.576

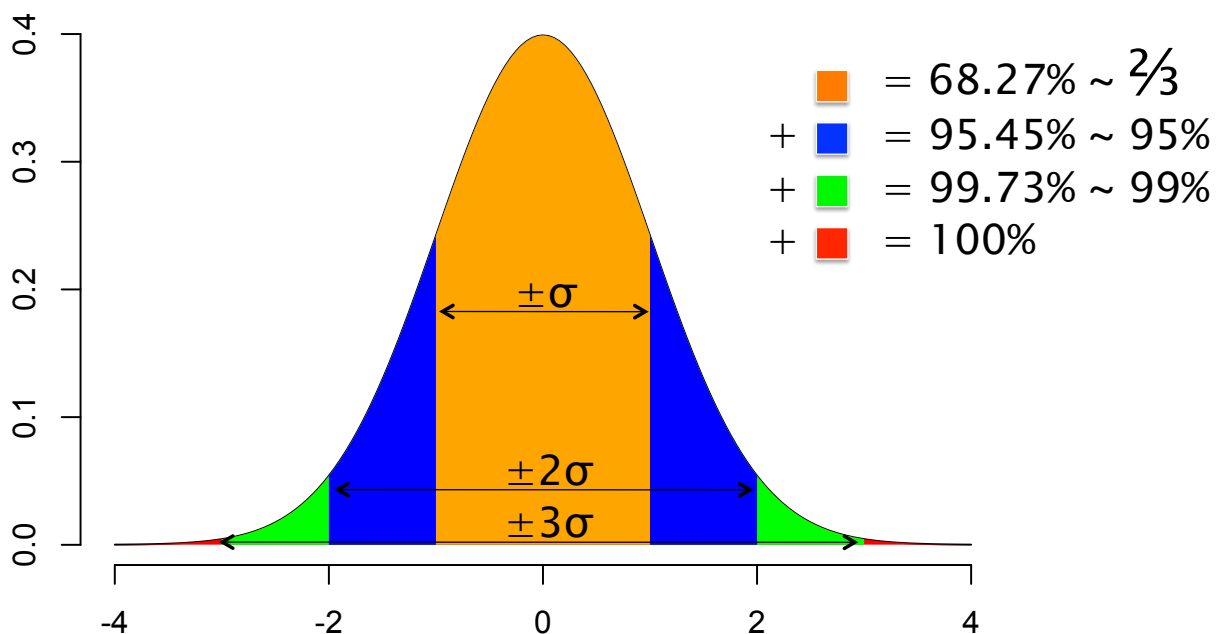
# Empirical rule

---

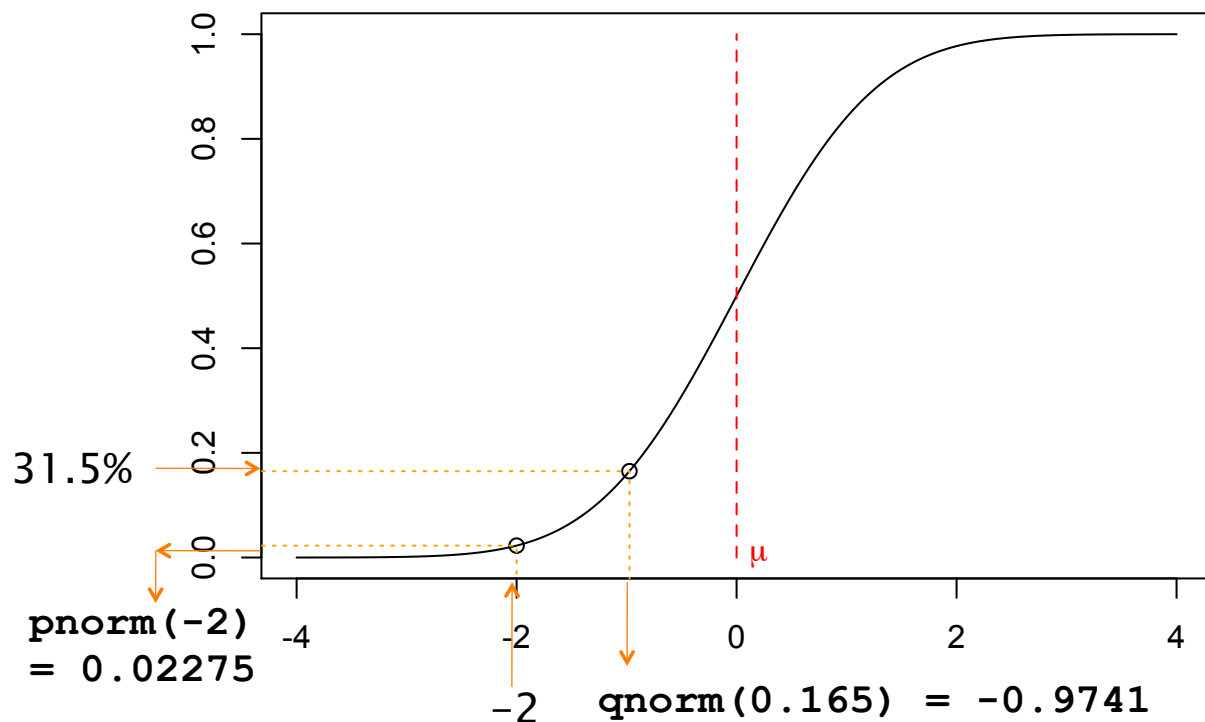
- Almost all the data fall within 3 standard deviations ( $Z=3$ ) of the mean
- 95% of the data fall within 2 standard deviations ( $Z=2$ ) of the mean
- Two thirds of the data falls within one standard deviation ( $Z=1$ ) of the mean

# Empirical rule

---



# Empirical rule and CDF



## Distributions in R

- `ddistr(x, size, prob, log=F)`
- `pdistr(q, size, prob, lower.tail=T, log.p=F)`
- `qdistr(p, size, prob, lower.tail=T, log.p=F)`
- `rdistr(n, size, prob)`

`binom` Binomial

`pois` Poisson

`unif` Uniform

`exp` Exponential

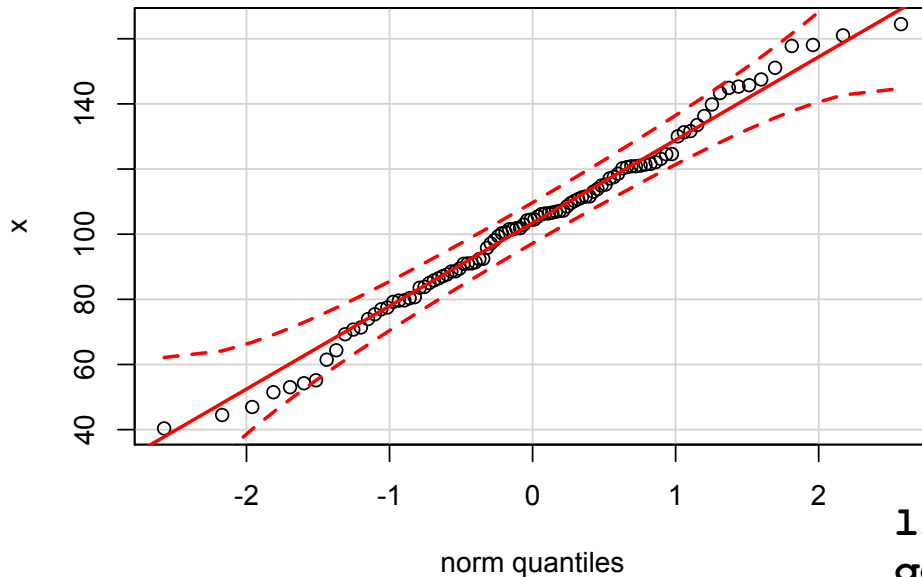
`norm` Normal

`t` Student t

# Checking for normality

---

- Quantile–Quantile plot



# Checking for normality

---

- Shapiro–Wilk test

$H_0$ : sample drawn from normal population

```
> shapiro.test(x)
```

Shapiro–Wilk normality test

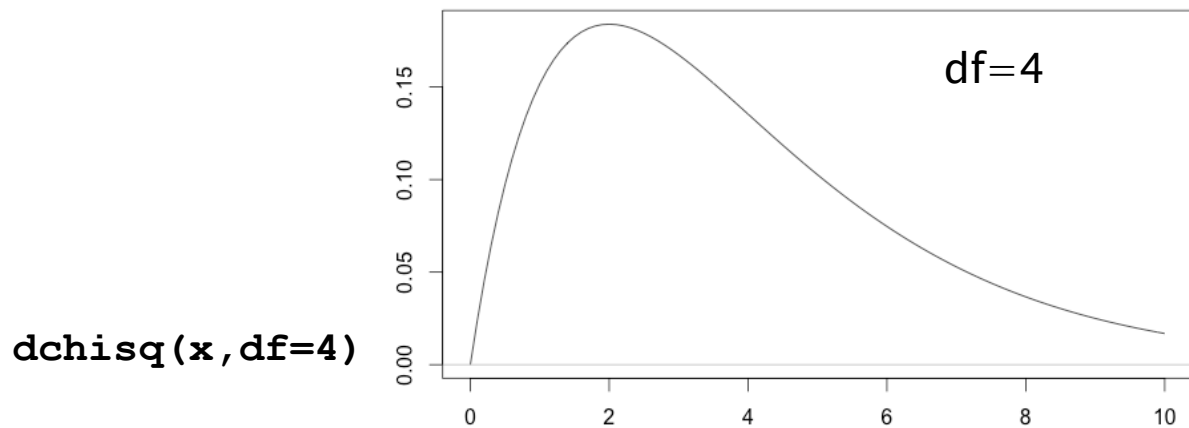
data: x

W = 0.9893, p-value = 0.6092

# Chi squared distribution

---

- Distribution of the sum of the square of  $k$  independent normal variables
  - ♦ Degrees of freedom  $df = k$



---

## CENTRAL LIMIT THEOREM



# Central Limit Theorem

---

- The random variable  $S_n$  that is constructed as the mean of a sample of size  $n$  of independent and identically distributed random variables

$$S_n = \frac{1}{n} \sum X_i$$

- For *large* values of  $n$ , the distribution of  $S_n$  is approximately normal with
  - ♦ Mean:  $\mu_S = \mu_X$
  - ♦ Standard deviation:  $\sigma_S = \frac{\sigma_X}{\sqrt{n}}$

## CLT

---

- The result is independent of the actual distribution of the original random variables
- The result is (mostly) independent of the size of the population
- The “precision” of the mean estimate depends on the sample size

# CLT – normalized form

---

- We can normalize  $S_n$ :

$$S_n^* = \frac{S_n - \mu}{\sigma/\sqrt{n}}$$

- The distribution for  $S_n^*$  is the standardized normal distribution

# CLT and Confidence Interval

---

- $1-\alpha$  confidence interval
  - ♦ Range where with  $P=1-\alpha$  lies  $\mu_X$
  - ♦  $P(|\mu_X - \bar{s}| < e) \geq 1-\alpha$
  - ♦ Typically expressed as:  $\bar{s} \pm e$
- What is the required sample size ( $n$ ) to achieve a fixed  $e$ ?
- CLT allows using the empirical rule
  - ♦ E.g. for the 95% CI:  $\bar{s} \pm 2\sigma_S$
  - ♦ Therefore 
$$e = 2\sigma_S = 2\frac{\sigma_X}{\sqrt{n}}$$

# Sample size and CLT

---

- Sampling voters in favor / against a given option
  - ♦  $p$  = real proportion of voters in favor
  - ♦ Items (0,1) are binomially distributed
  - ♦  $\mu_X = 1 \cdot p + 0 \cdot (1-p) = p$
  - ♦  $\sigma_X = \sqrt{p \cdot (1-p)} \leq \sqrt{1/4} = 1/2$

# Sample size and CLT

---

- The distribution of  $\bar{s}$  is approximately normal
- A 95% interval is:  $\bar{s} \pm 2\sigma_S$
- Therefore:

$$e > 2\sigma_S = 2 \frac{\sigma_X}{\sqrt{n}} > 2 \frac{1/2}{\sqrt{n}} = \frac{1}{\sqrt{n}}$$

Binomial

- ♦ That is:  $n_{95\%CI} > \frac{1}{e^2}$        $n_{99\%CI} > \frac{9}{4e^2}$

# Sample size and CLT

---

- In practice:
  - ♦ Fix the confidence interval
  - ♦ Find the required minimum sample size

	95% CI	99% CI
±5%	400	900
±3%	1,111	2,500
±1%	10,000	22,500

## Student's t distribution

---

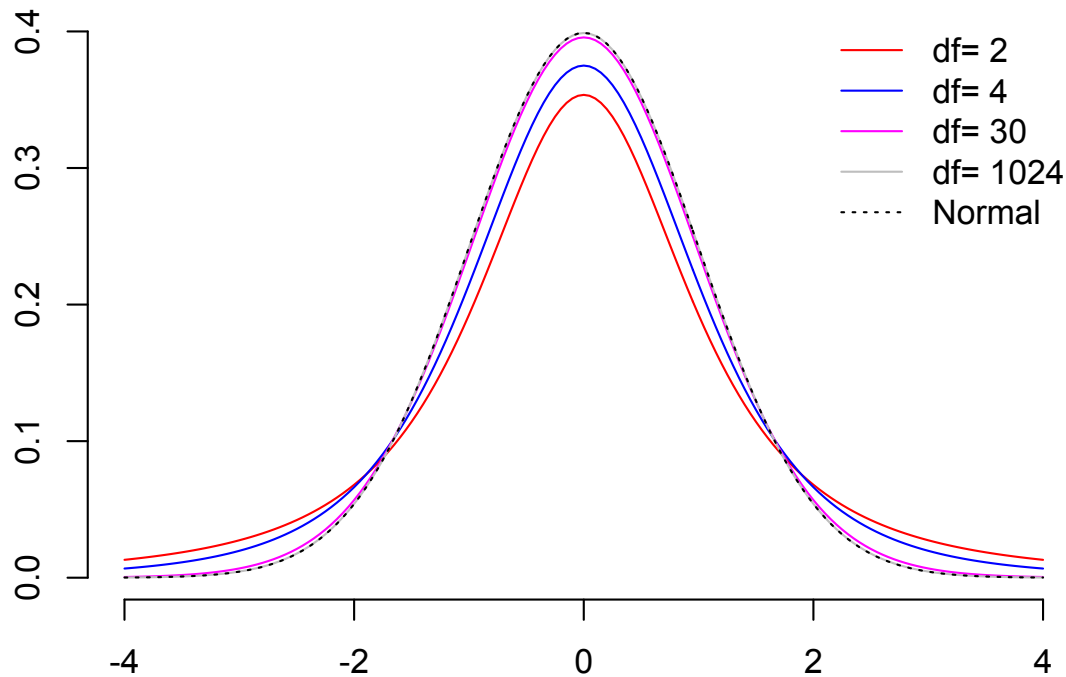
- Given a sample  $\{x_1 \dots x_i \dots\}$  of size  $n$  from a **normally distributed** population, with mean  $\bar{x}$  and standard deviation  $s$
- The t value is defined as:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

- Student's t distribution with  $n-1$  degrees of freedom describe the distribution of t

# Student's t distribution

---



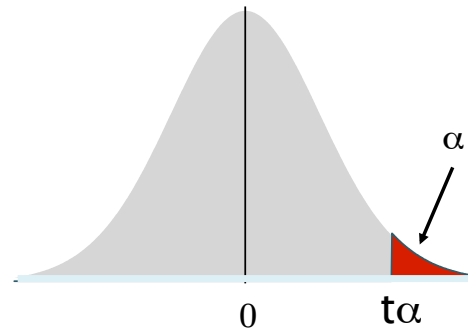
# Student's t distribution

---

- Complements the CLT when
  - ♦ Small sample sizes
  - ♦ But drawn from a **normal** distribution
- For large sample sizes tends to a normal distribution

# Table of Critical Values of $t$

df	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
$\infty$	1.282	1.645	1.960	2.327	2.576



With  $df = 24$  and  $\alpha = 0.05$ ,  $t_\alpha = 1.711$ .

## Finite population

- CLT valid for infinite populations
  - ◆ Holds well for populations large w.r.t. sample size
- If the population is small, use the Finite Population Correction factor

$$fpc = \sqrt{\frac{N - n}{N - 1}}$$

$N$  = population size  
 $n$  = sample size

# Sampling

---

- Population
  - ◆ The universe of entities
    - People, objects or any item
- Sample
  - ◆ A subset of the population
- Census
  - ◆ Data from the entire population

## Parameter vs statistic

---

- Parameter: feature of population
  - ◆  $\mu$ : mean
  - ◆  $\sigma$ : standard deviation
- Statistic: feature of the sample
  - ◆  $\bar{x}$ : mean
  - ◆  $s$ : standard deviation

# Parameters and statistics

---

- Mean

$$\mu \sim \bar{x} = \frac{1}{n} \sum_{i=0}^n x_i$$

- Standard deviation

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=0}^n (x_i - \bar{x})^2} \quad s = \sqrt{\frac{1}{N} \sum_{i=0}^n (x_i - \bar{x})^2}$$

---

## HYPOTHESIS TESTING



# Types of Hypotheses

---

- Research Hypothesis
  - ♦ a statement of what the researcher believes will be the outcome of an experiment or a study.
- Statistical Hypotheses
  - ♦ a more formal structure derived from the research hypothesis.
- Substantive Hypotheses
  - ♦ a statistically significant difference does not imply or mean a material, substantive difference.

## Statistical Hypothesis testing

---

Assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed?

# Steps

---

- H** { 1. Establish hypotheses  
♦ state the null and alternative hypotheses.
- T** { 2. Determine the appropriate statistical test and sampling distribution.  
3. Specify the Type I error rate ( $\alpha$ ).  
4. State the decision rule.  
5. Gather sample data.
- A** { 6. Calculate the value of the test statistic.  
7. State the statistical conclusion.
- B** { 8. Make a managerial decision.

## One sample

---

- $H_0: \mu = 0$
- Given a sample of  $n$  elements
  - ♦ Be  $s$  the sample standard deviation
  - ♦ The  $t$  value is distributed according the the Student's  $t$  distribution with  $n-1$  degrees of freedom

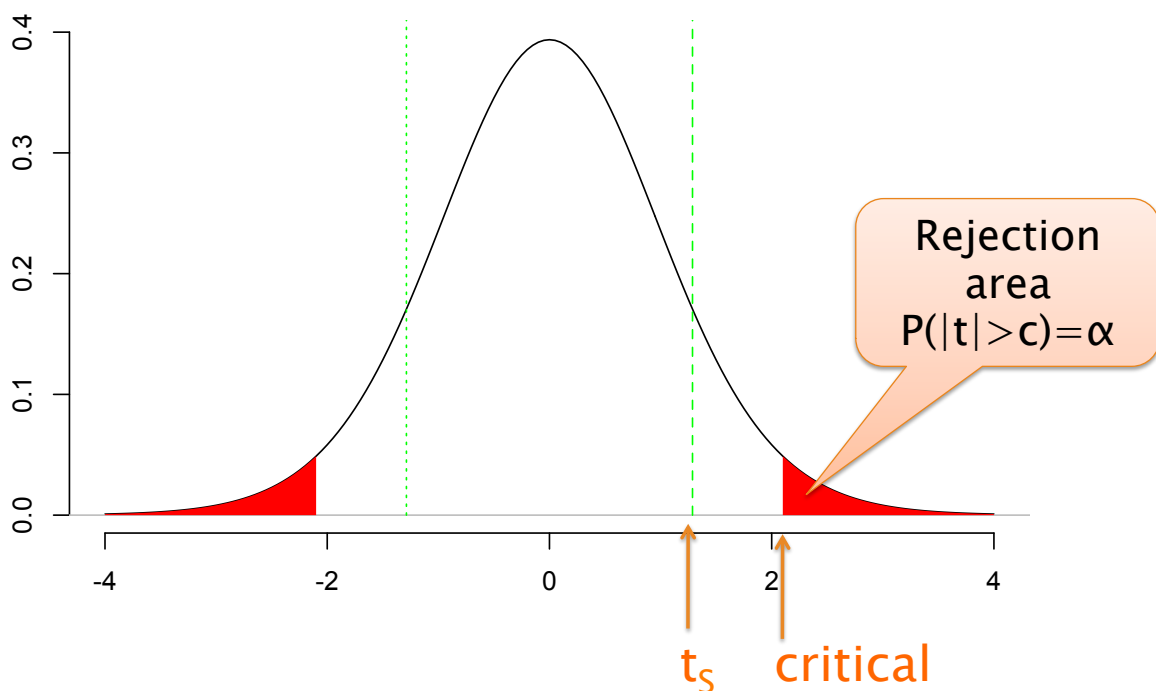
# One Sample – Example

- Samples  $n = 20$
- $\bar{x} = 0.473$
- $s = 1.645$

$$t_s = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{0.473 - 0}{1.645/\sqrt{20}} = 1.287$$

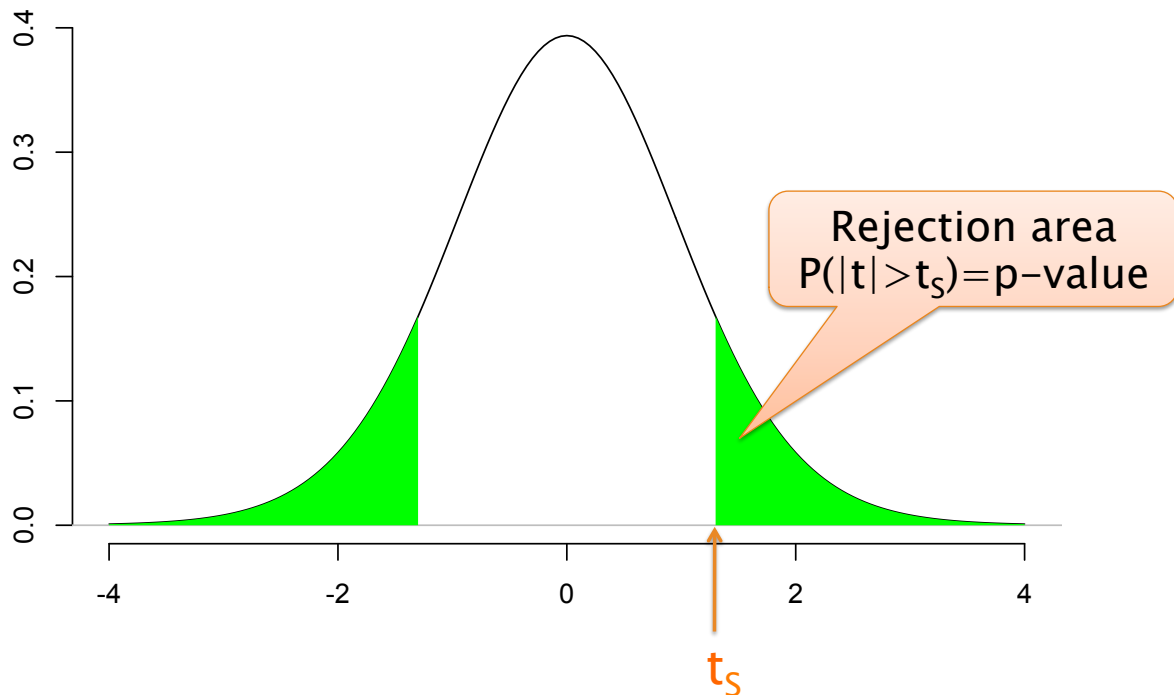
- Critical value c.r =  $\pm 2.093$ 
  - ♦  $qt(1-\alpha/2, n-1)$

# One Sample – Example



# One sample – Example

---



# One sample – Example

---

- Rejection decision criteria:
  - ♦  $t_s >$  critical value
    - $1.287 > 2.093 \rightarrow$  fail to reject
  - ♦  $p\text{-value} < \alpha$ 
    - $0.213 < 0.05 \rightarrow$  fail to reject
- **t.test(samples, mu=0)**
  - ♦  $t = 1.2873, df = 19, p\text{-value} = 0.2134$

# One sample – Example

---

- Confidence interval
  - ♦ The theoretical range of  $\mu$  within which  $H_0$  cannot be rejected

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} = \pm \text{critical}$$

$$\frac{0.473 - \mu}{1.645/\sqrt{20}} = \pm 2.093$$

$$\mu = 0.473 \pm \frac{2.093 \cdot 1.645}{\sqrt{20}} = [-0.297; 1.243]$$

1- $\alpha$  CI

# One sample

---

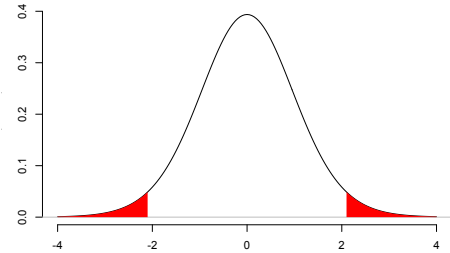
- $H_0: \mu = \mu_0$
- Fail to reject  $H_0 \Leftrightarrow \mu_0 \in \text{CI}$
- Reject  $H_0 \Leftrightarrow \mu_0 \notin \text{CI}$

# One- vs. Two-tailed tests

---

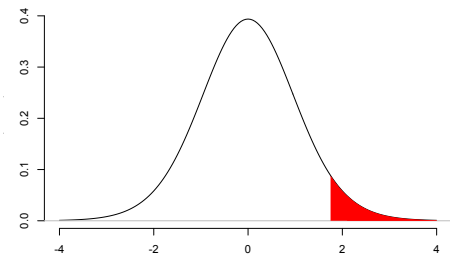
- Un-directional hypothesis

- ♦  $H_0: \mu = 0$
- ♦  $H_a: \mu \neq 0$



- Directional hypothesis

- ♦  $H_0: \mu \leq 0$
- ♦  $H_a: \mu > 0$



# One- vs. Two-tailed tests

---

- One-tailed test makes it "easier" to reject the null hypothesis
  - ♦ The critical area is larger
  - ♦ The critical value is closer to the mean
  - ♦ The p-values is divided by two

# One-tailed or two-tailed?

---

- If  $H_a$  simply says the two means will be different, but doesn't predict a direction to the difference, then you would use the two-tailed t-test value for comparison with the critical value
- If  $H_a$  predicts a difference in a particular direction (one mean will be larger than the other), then you would use a one-tailed t-test

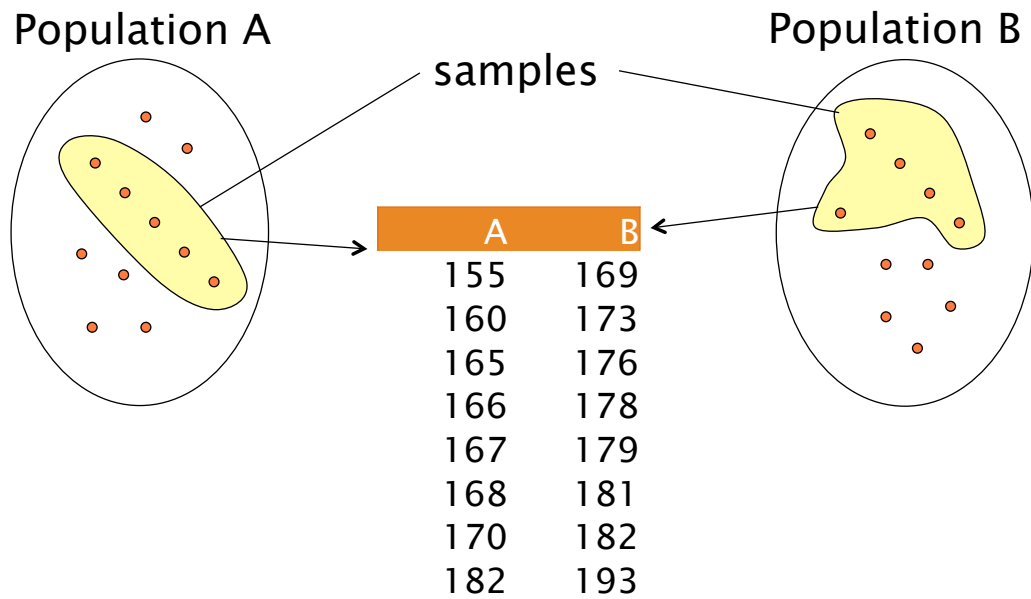
# Two-samples

---

- Often the hypotheses compare measures from two different samples
  - ♦ Typically two levels of the main factor
    - i.e. with and without the treatment

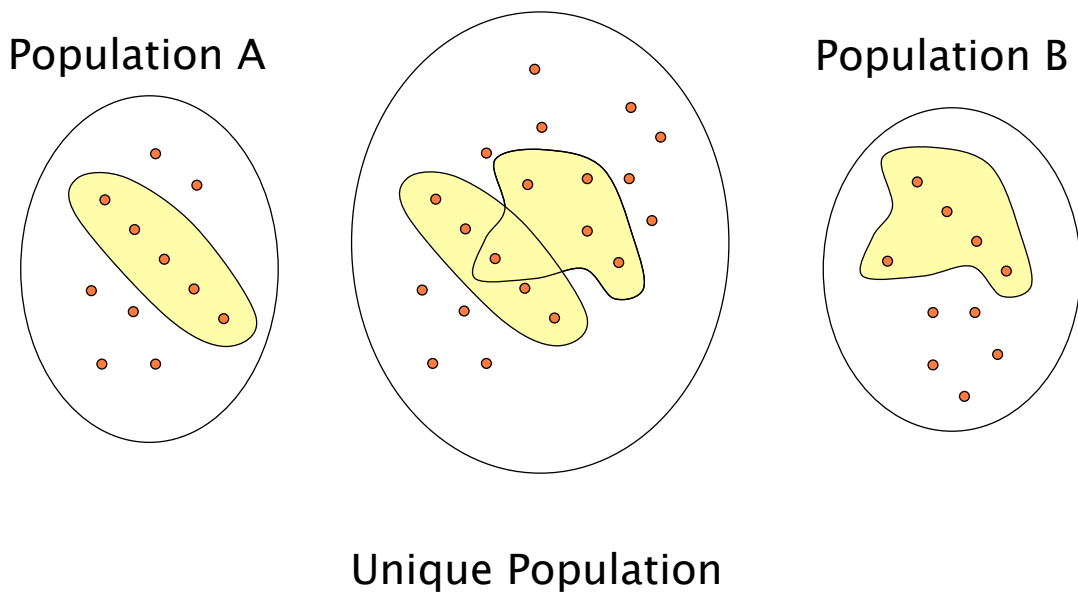
# Sampling

---



# Sampling

---





# Two sample t-test

---

- `t.test( response ~ factor,  
          data=d)`
  - ♦ `alternative =`
    - `c( "two.sided", "less", "greater"),`
  - ♦ `mu = 0`

---

## NONPARAMETRIC TESTS

# Parametric vs. Nonparametric

---

- Parametric tests are based on assumptions:
  - ♦ data being analyzed are randomly selected from a **normally distributed population**.
  - ♦ quantitative measurement that yield interval or ratio level data.
- Nonparametric tests require fewer assumptions
  - ♦ Sometimes called “distribution-free” statistics.
  - ♦ Techniques available for use with nominal or ordinal data.

## Pros of Nonparametric

---

- Sometimes there is no parametric alternative to the use of nonparametric statistics.
- Can be used to analyze
  - ♦ nominal data.
  - ♦ ordinal data.
- Computation less complicated than parametric statistics
  - ♦ particularly for small samples.
- Probability statements obtained from most nonparametric tests are exact probabilities.

# Cons of Nonparametric

---

- Can be wasteful of data if parametric tests are available
- Not as widely available and well known as parametric tests.
- For large samples, the calculations for many nonparametric statistics can be tedious.

# Wilcoxon signed rank test

---

- Nonparametric alternative to one sample t test
  - ♦ Paired difference test
  - ♦ Also one sample test
- Procedure
  - ♦  $R_i = \text{Rank } |X_i - \theta|$
  - ♦ Sum the ranks of positive  $X_i - \theta$

$$W_+ = \sum_{i=1}^n \phi_i R_i \quad \phi_i = \begin{cases} 0 & \text{if } X_i - \theta < 0 \\ 1 & \text{if } X_i - \theta > 0 \end{cases}$$

# Wilcoxon signed rank test

---

- Expected value if  $\theta$  is median:
  - ♦  $W = \frac{n(n+1)}{4}$
- One sample:
  - ♦ `wilcox.test(x)`
- Two samples:
  - ♦ `wilcox.test(x,y,paired=T)`

# Wilcoxon signed rank test

---

- Nonparametric alternative to one sample t test
  - ♦  $H_0$ : median =  $\theta$
- One sample:
  - ♦ `wilcox.test(x)`
- Two samples (paired differences):
  - ♦ `wilcox.test(x,y,paired=T)`

# Wilcoxon signed rank test

---

- Procedure

- ♦ Remove all  $X_i = \theta$
  - ♦  $R_i = \text{Rank } |X_i - \theta|$
  - ♦  $\phi_i = \text{sign}(X_i - \theta)$
- $$W = \sum_{i=1}^N \phi_i \cdot R_i$$

- $W$  is normally distributed (for large  $N$ )

$$z = \frac{W - 0.5}{\sigma_W}, \sigma_W = \sqrt{\frac{N(N+1)(2N+1)}{6}}$$

# Mann-Whitney U Test

---

- Nonparametric counterpart of the t test for independent samples
- Does not require normally distributed populations
- Assumptions
  - ♦ Independent Samples
  - ♦ At Least **Ordinal** Data

# Mann–Whitney U Test: samples

---

- Size of samples:  $n_1, n_2$
- If both  $n_1$  and  $n_2$  are  $\leq 10$ , the small sample procedure is appropriate.
- If either  $n_1$  or  $n_2$  is greater than 10, the large sample procedure is appropriate.

## MW Test

---

- Small samples:
  - ♦ For each item in sample  $a$  count how many items of sample  $b$  have lower rank
  - ♦ Sum all the values
- Large samples: minimum  $U_s$

$$U_s = \sum R_{si} - \frac{n_s(n_s + 1)}{2}$$

# MW Test

- Use the smallest U for computation
- Normal approximation for  $N > 20$

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sigma_U}, \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

- For small N tables are available

## MW Test

- Assign a rank to the union of the samples
  - ♦  $a = (4, 7, 3, 5)$
  - ♦  $b = (6, 9, 12, 10)$

sample	values	rank
a	3	1
a	4	2
a	5	3
b	6	4
a	7	5
b	9	6
b	10	7
b	12	8

$$U_a = (1 + 2 + 3 + 5) - \frac{4 \cdot (4 + 1)}{2} = 1$$

$$U_b = (4 + 6 + 7 + 8) - \frac{4 \cdot (4 + 1)}{2} = 15$$

$$z = \frac{1 - \frac{4 \cdot 4}{2}}{\sqrt{\frac{4 \cdot 4 \cdot (4 + 4 + 1)}{12}}} = \frac{7}{\sqrt{12}} = -2.02$$

p-value = 0.043

# Nominal metrics tests

---

- Pearson Chi Squared test
  - ♦ Independence of nominal variables
    - Contingency table
  - ♦ Difference in distribution frequencies
    - Table with paired distributions (?bind)
  - ♦ Goodness of fit
- Fisher exact test
  - ♦ In 2x2 or 3x3 cases
  - ♦ Small samples

# Pearson Chi Squared test

---

- Comparing observed frequencies to expected ones
- Test statistic: 
$$X^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$
  - ♦ Asymptotically approaches a  $\chi^2$  distribution with  $n-p$  degrees of freedom
  - ♦  $p$  = number of parameters - 1



# Chi-squared test

		Variable 2			Marginals
		L	M	H	
Variable 1	L	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$m_{1,*}$
	M	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$m_{2,*}$
	H	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	$m_{3,*}$
		$m_{*,1}$	$m_{*,2}$	$m_{*,3}$	N

## Pearson chi squared test

- Expected values in case of independent variables

$$E_{i,j} = \frac{m_{i,*} \cdot m_{*,j}}{N}$$

- # parameters:  $c + r$
- Degrees of freedom  $df = N - (c + r - 1)$ 
  - Since  $N = c \cdot r$ ,  $df = (c - 1)(r - 1)$

- `chisq.test(t)`

# 2x2 Contingency

		Outcome		
		Correct	Wrong	
Treatment	+	$f_{+,C}$	$f_{+,W}$	$n_+$
	-	$f_{-,C}$	$f_{-,W}$	$n_-$
		$n_C$	$n_W$	$N$

## Odds vs. Proportions

- Proportions

- Correct with treatment +:  $\frac{f_{+,C}}{f_{+,C} + f_{+,W}}$
- Correct with treatment -:  $\frac{f_{-,C}}{f_{-,C} + f_{-,W}}$

- Odds

- Correct vs. wrong with treatment +:  $\frac{f_{+,C}}{f_{+,W}}$
- Correct vs. wrong with treatment -:  $\frac{f_{-,C}}{f_{-,W}}$

- Odds ratio:  $\frac{f_{+,C} \cdot f_{-,W}}{f_{+,W} \cdot f_{-,C}}$

# Fisher exact test

---

- $H_0: OR = 1$ 
  - ♦ p-value: probability of observing at least as such an extreme OR given the observed marginals
- `fisher.test(t)`

# Multiple comparisons

---

- As the number of comparisons increases, it becomes more likely that the groups being compared will appear to differ in terms of at least one attribute.
- Family-wise error rate
  - ♦  $\alpha_{FW} = 1 - (1 - \alpha_c)^n$
- Bonferroni correction
  - ♦ Setting:  $\alpha_c = \alpha/n$
  - ♦ Implies:  $\alpha_{FW} \leq \alpha$

---

# ANALYSIS OF VARIANCE

## Purpose

---

- Compare more than two populations (instead of just two populations as done with the t-test)
  - ◆ Several possible levels for the main factor
    - One-way
  - ◆ Used when to analyze the effect of different factors on the dependent variable
    - N-way

# Assumptions

---

- Observations are drawn from normally distributed populations
- Observations represent random samples from the populations
- Homoscedasticity: variances of the populations are equal

# Sum of squares

---

- Random effect model

- ♦  $Y_{ij} = \mu + B_i + W_{ij}$

- Sum of squares  $SST = \sum_i \sum_j (Y_{ij} - \bar{Y})^2$

$$SST = \underbrace{\sum_j n_j (\bar{Y}_j - \bar{Y})^2}_{SSC} + \underbrace{\sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2}_{SSE}$$

Treatment SS  
Between treatments SS

Error SS  
Within treatments SS

# ANOVA

---

- Between treatment variance:
  - ♦  $MSC = SSC/df_C$ 
    - $df_C = \text{\#levels} - 1$
- Within treatment variance:
  - ♦  $MSE = SSE/df_E$ 
    - $df_E = N - \text{\#levels}$
- $F = MSC / MSE$ 
  - ♦ F follows an F-distribution

# ANOVA

---

- AOV
  - ♦ `aov( output ~ factor, data=data)`
- ANOVA test
  - ♦ `summary(aov(output ~ factor,...))`

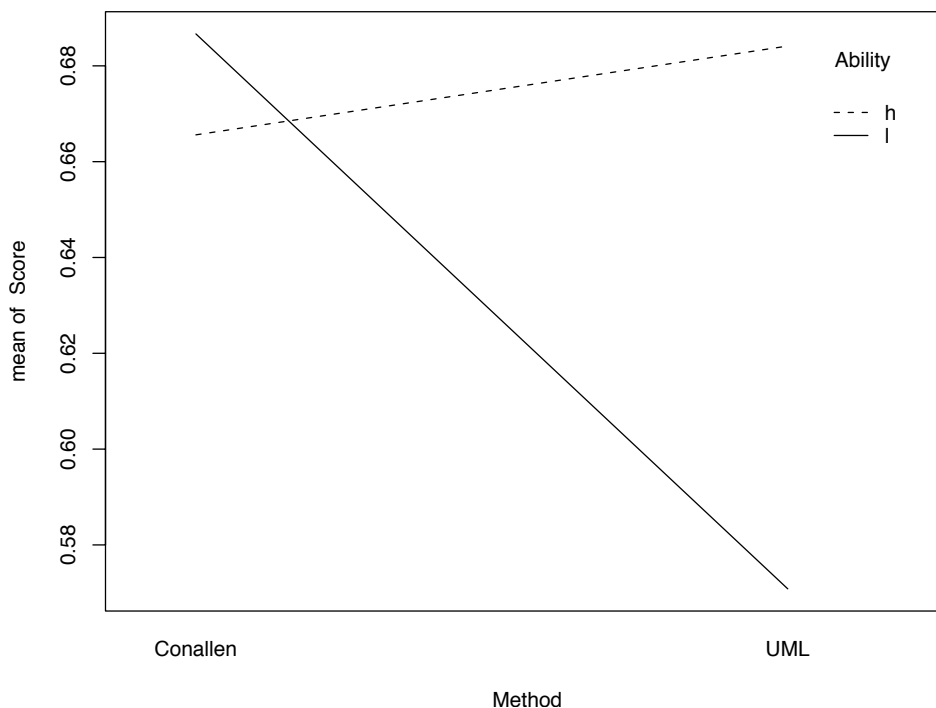
# Two-way ANOVA

---

- Blocked design
  - ♦  $SST = SSC + SSR + SSE$
  - ♦ `aov( output ~ factor + block, ...)`
  - ♦ Factor and block are independent
- Factorial design
  - ♦ `aov( output ~ factor1*factor2, ...)`
  - ♦ Factors are not independent
  - ♦ There may be interaction

## Interaction diagram

---



# References

---

- Kabacoff. “R in Action”, Manning, 2011
- Lots of online resources