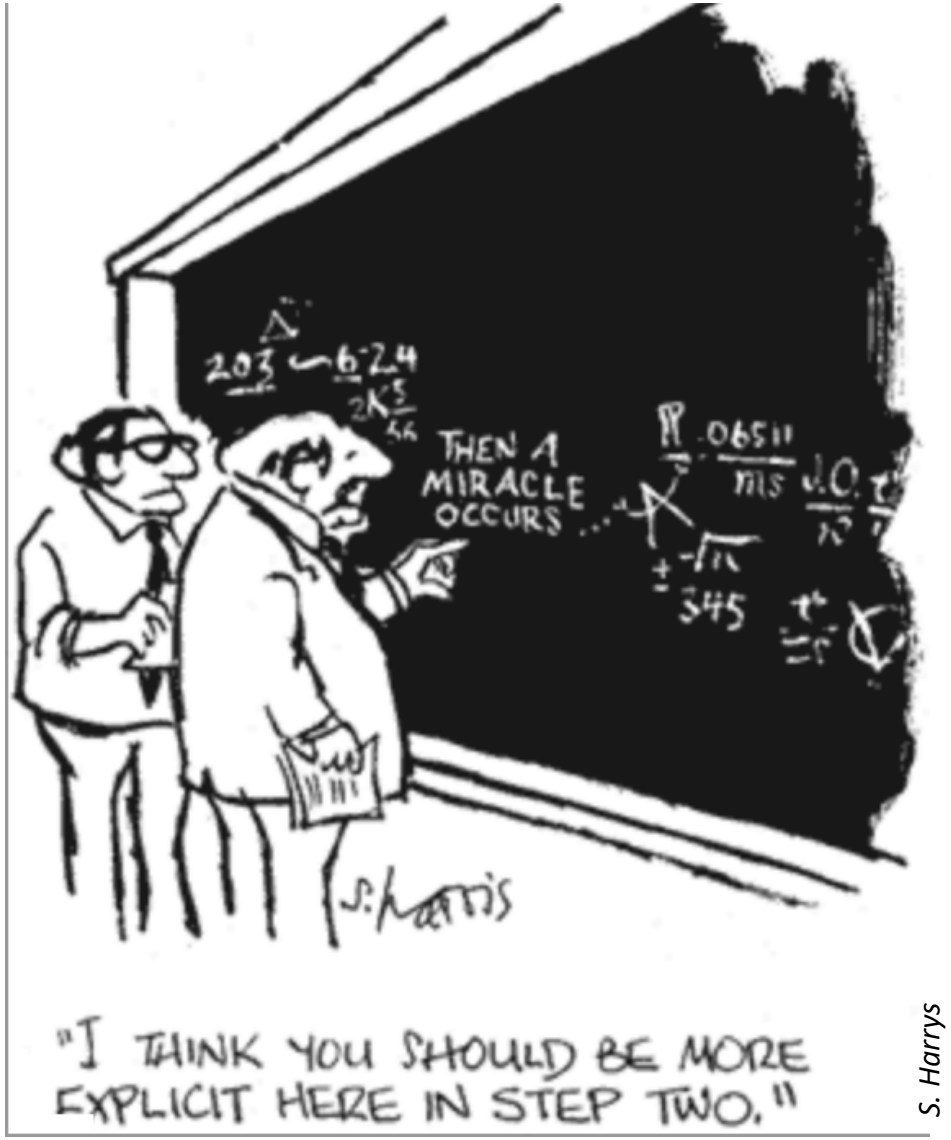
Data analysis in R applied to Marine Science

Day 3. BASIC STATISTICAL ANALYSIS
29 January 2019

Dr. Tamara Huete-Stauffer – Dr. Grégoire Michoud – Dr. Daffne López-Sandoval – Dr. Malika Kheireddine

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology
RSRC Red Sea Research Center

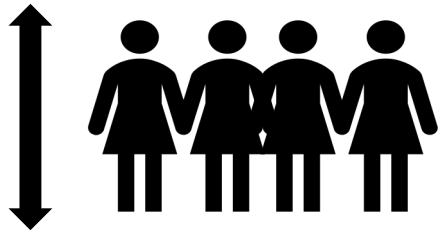
Revisiting the basics



Inferential statistics

Say we want to determine the average height of woman PhD and postdocs in the campus, and let's assume we have data from every single woman in the university (N)

Height all woman Population = N

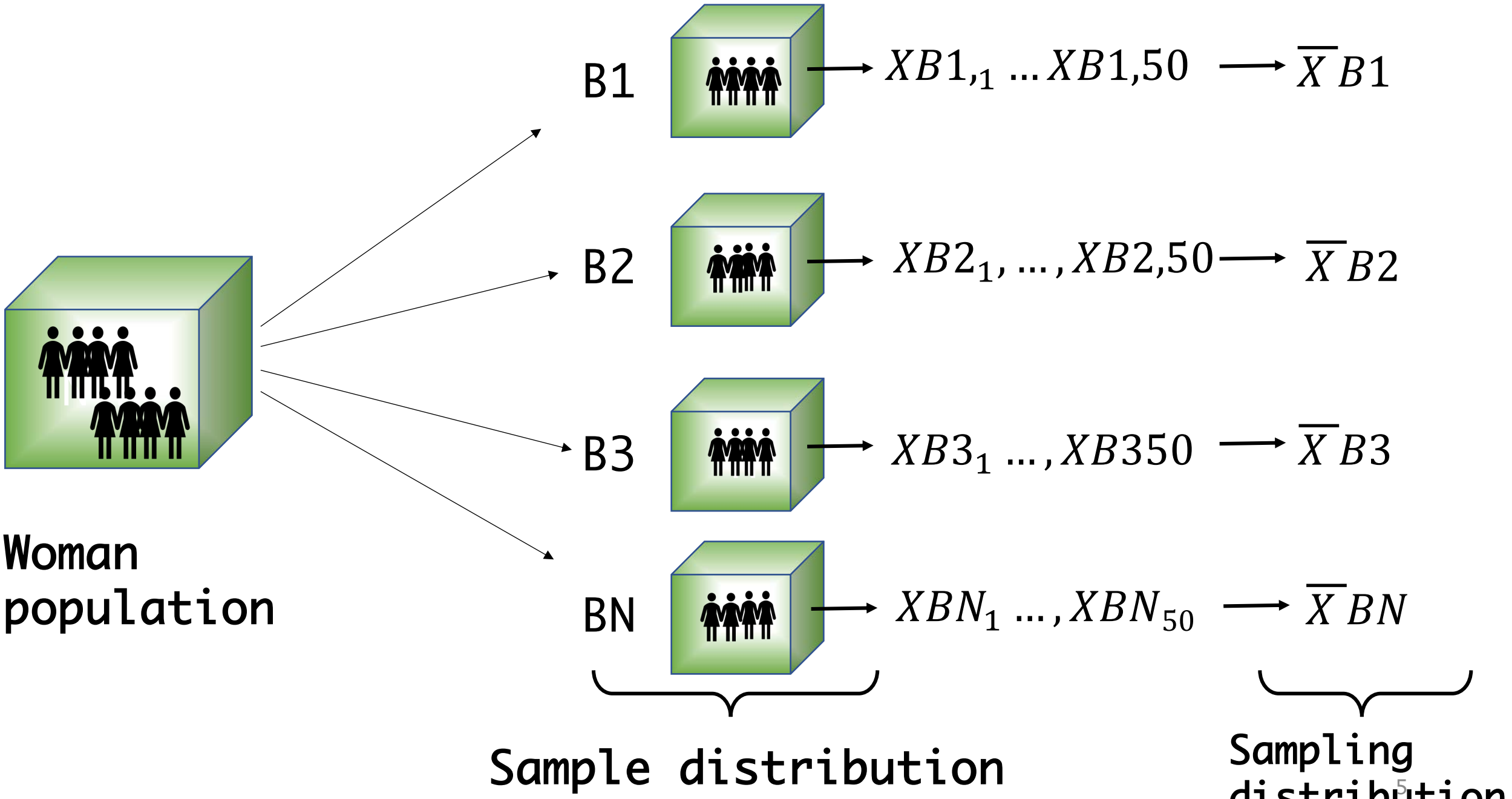


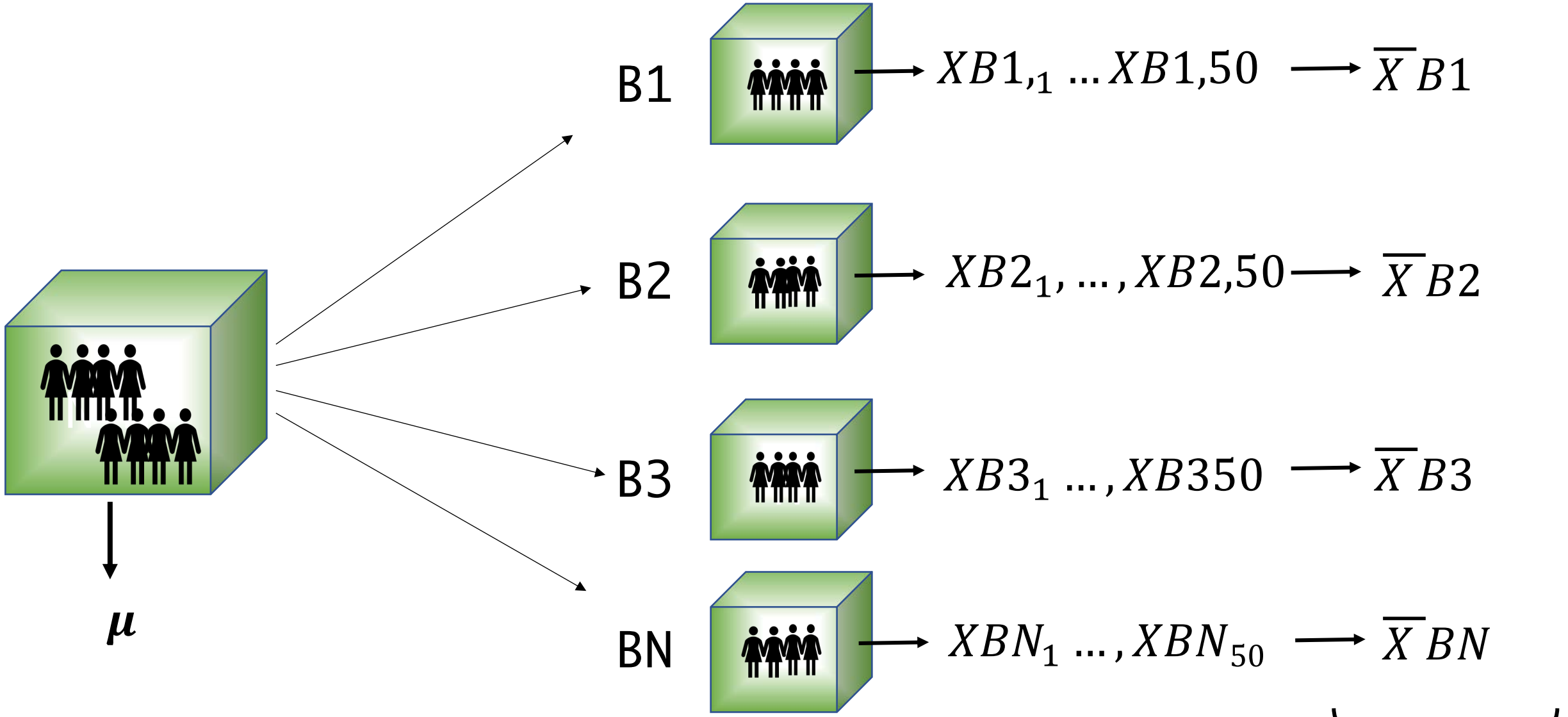
$$\mu = \frac{x_1 + x_2 \ldots + x_N}{N}$$ → Average

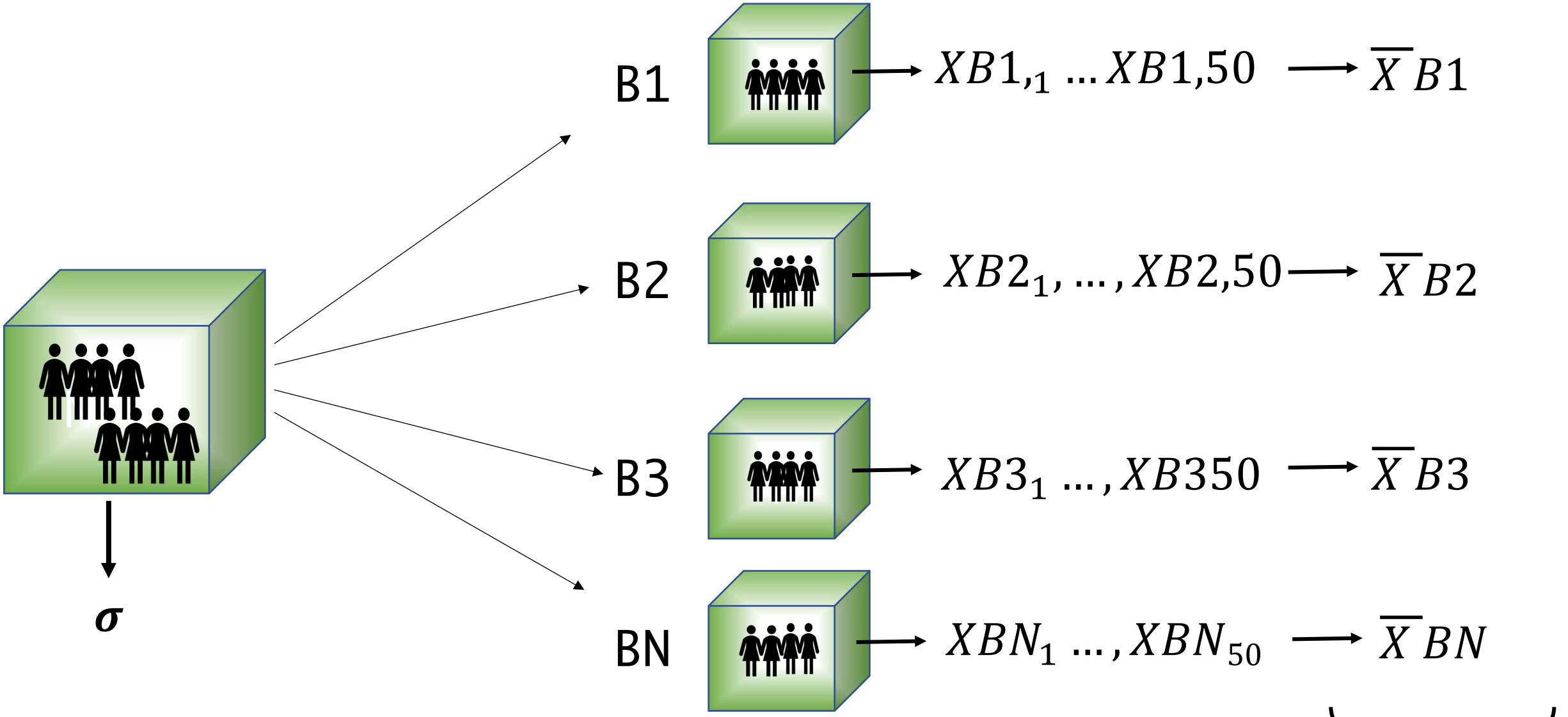
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N}}$$ → Variability

Inferential statistics

We don’t expect σ value be very small

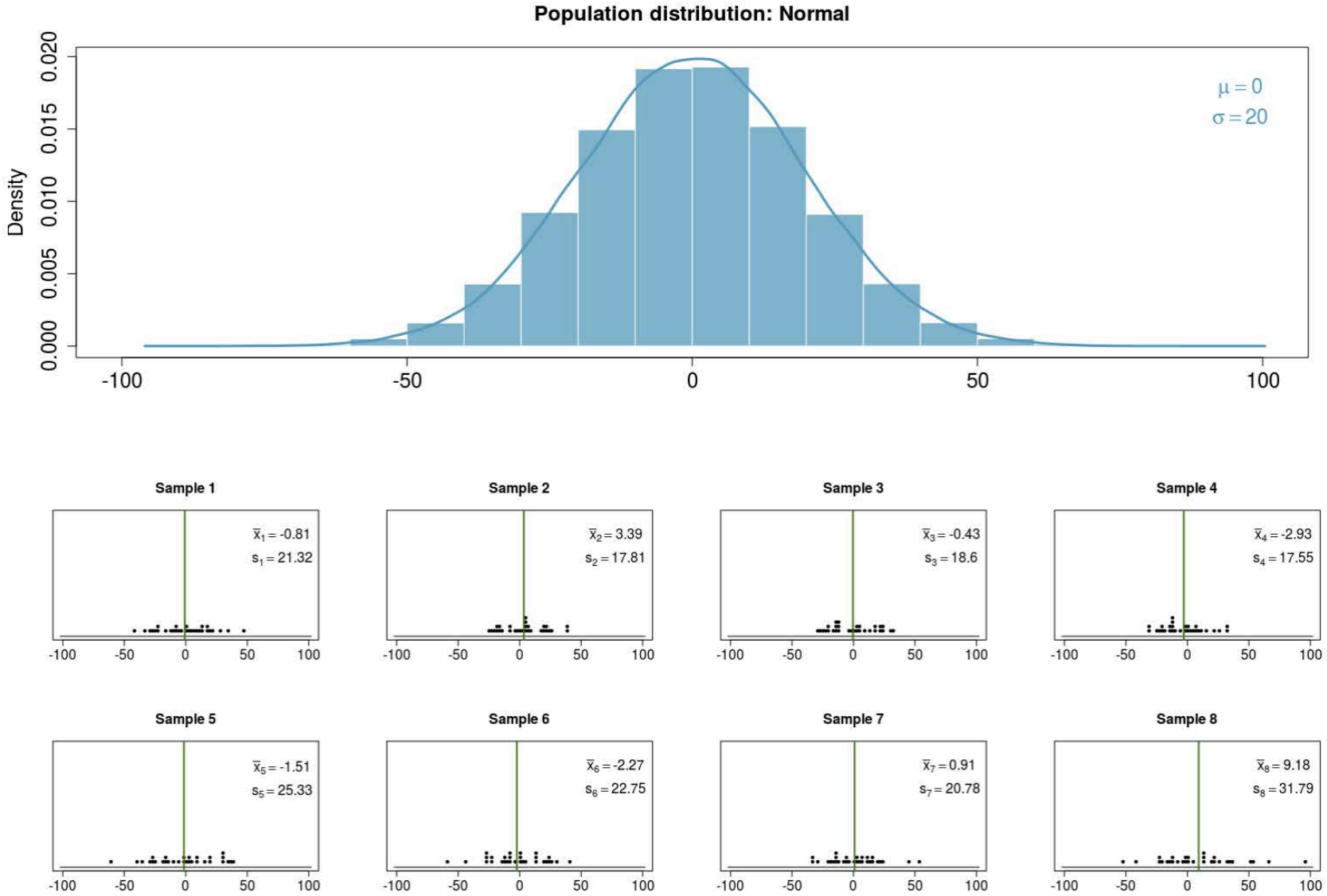

B1
$XB1_{,1} \ldots XB1{,}50 \longrightarrow \overline{X}\,B1$
B2
$XB2_1, \ldots, XB2{,}50 \longrightarrow \overline{X}\,B2$
B3
$XB3_1 \ldots, XB350 \longrightarrow \overline{X}\,B3$
BN
$XBN_1 \ldots, XBN_{50} \longrightarrow \overline{X}\,BN$
Woman population
Sample distribution
Sampling distribution


B1
$XB1_{,1} \ldots XB1{,}50 \longrightarrow \overline{X}\,B1$
B2
$XB2_1, \ldots, XB2{,}50 \longrightarrow \overline{X}\,B2$
B3
$XB3_1 \ldots, XB350 \longrightarrow \overline{X}\,B3$
BN
$XBN_1 \ldots, XBN_{50} \longrightarrow \overline{X}\,BN$
$\boldsymbol{\mu}$
Mean $\overline{\boldsymbol{X}} \approx \boldsymbol{\mu}$

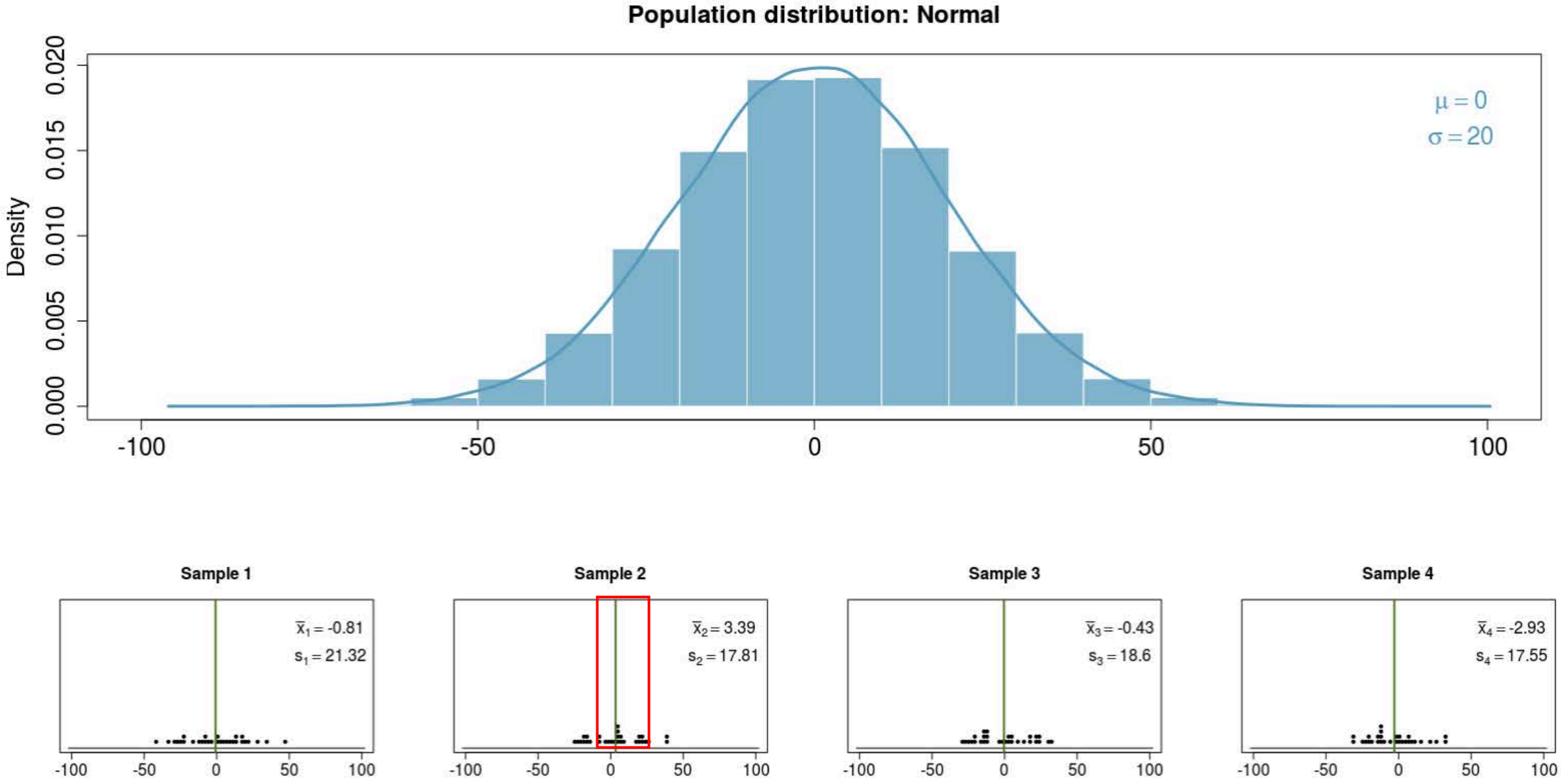


The standard deviation of the sample means (SE)

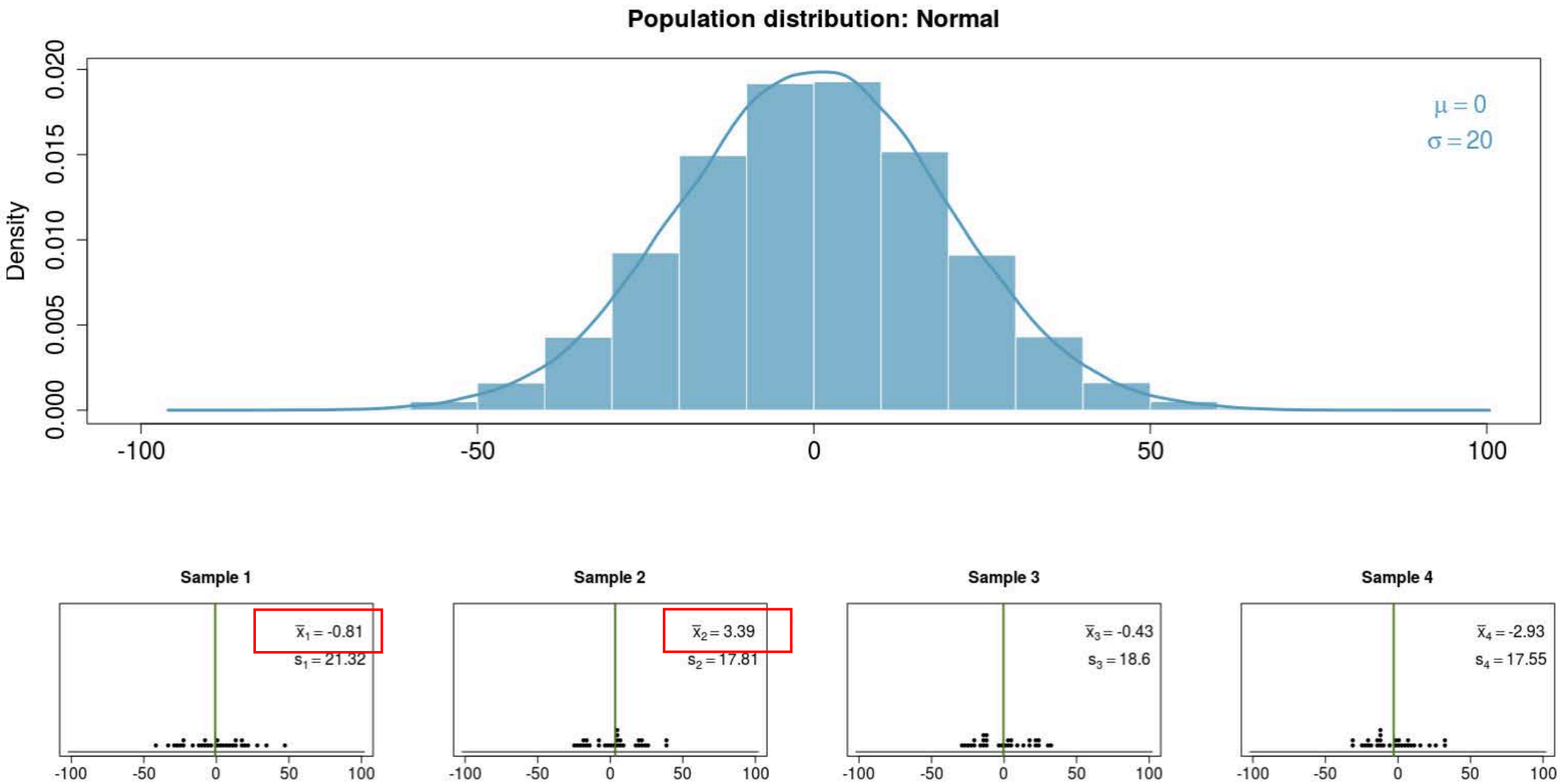
$\longrightarrow$ SE (SD $\bar{X} < \sigma$)



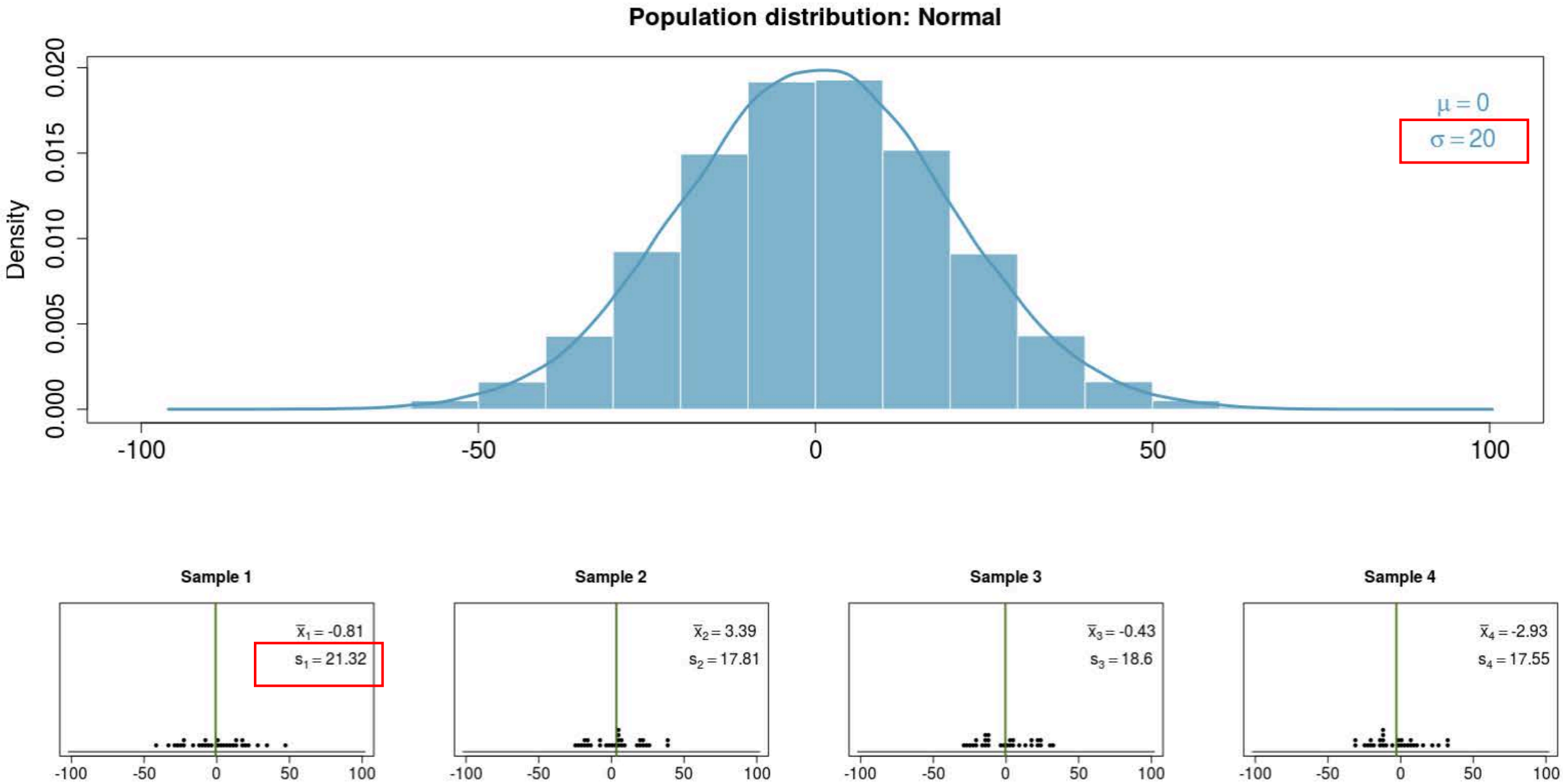
https://gallery.shinyapps.io/dist_calc/



https://gallery.shinyapps.io/dist_calc/



https://gallery.shinyapps.io/dist_calc/

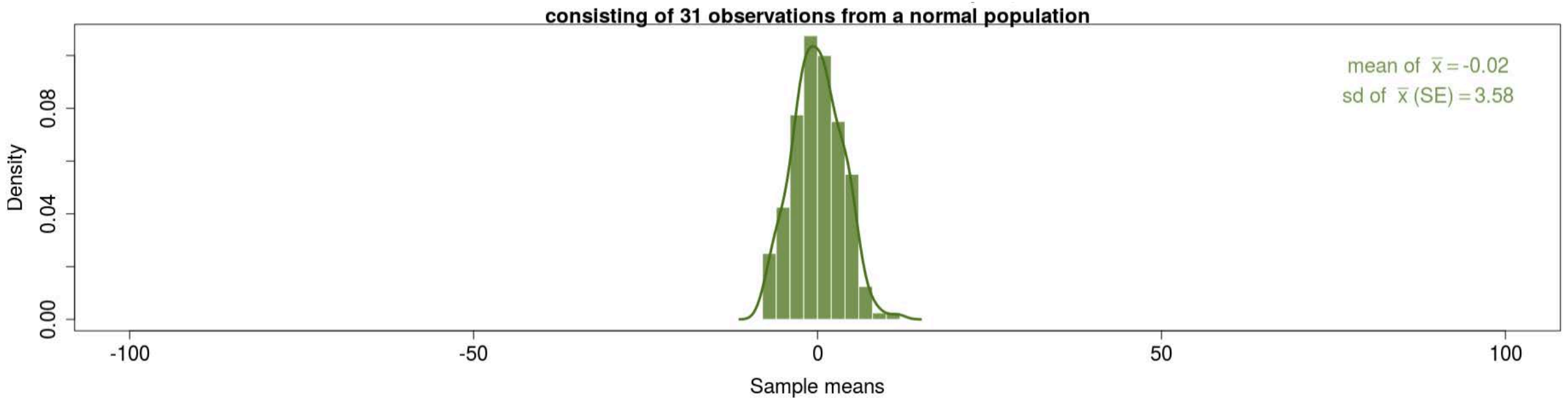


https://gallery.shinyapps.io/dist_calc/

Distribution of sample means

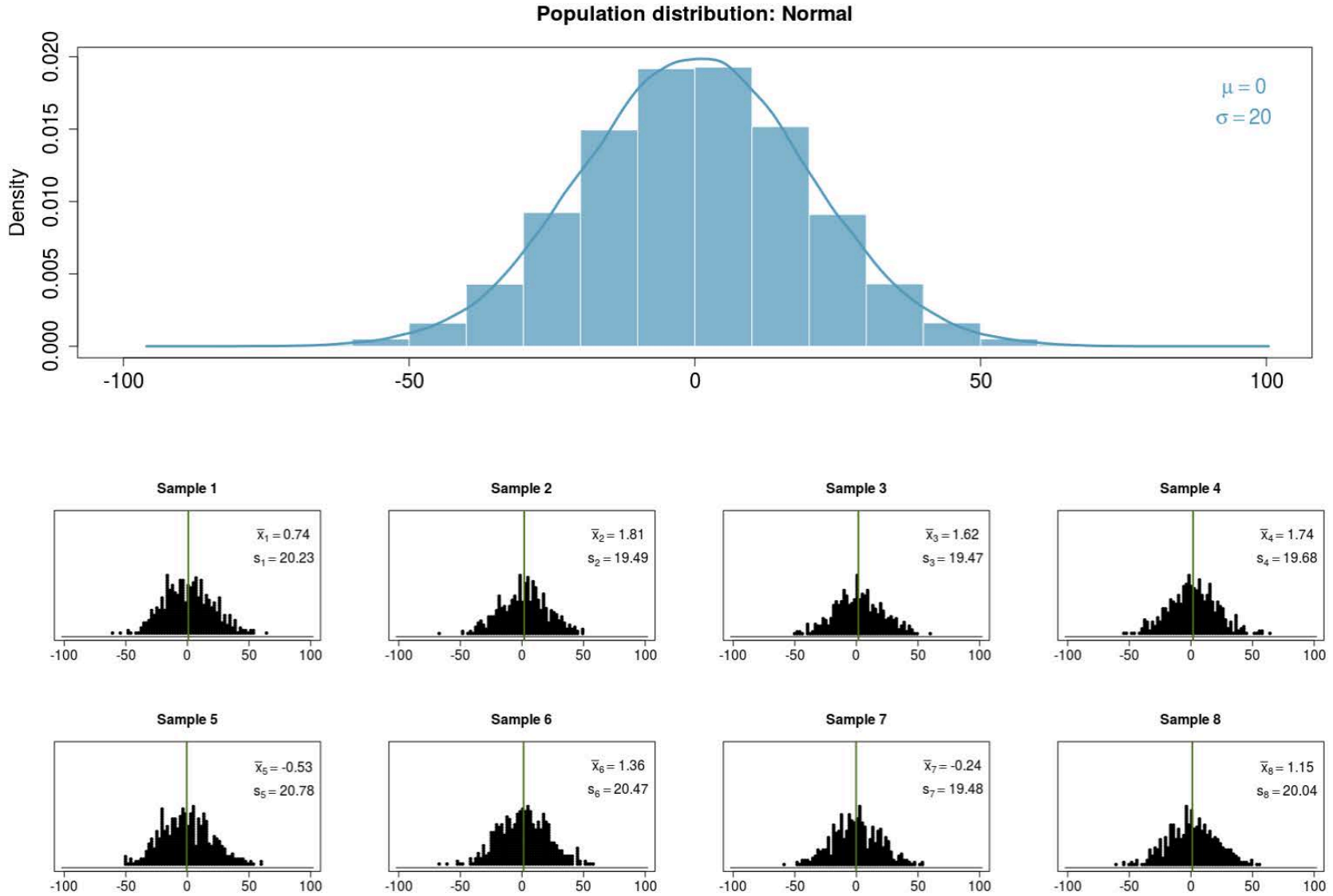
Distribution of the sample means or our sampling distribution

The sampling distribution illustrates how the variability looks like

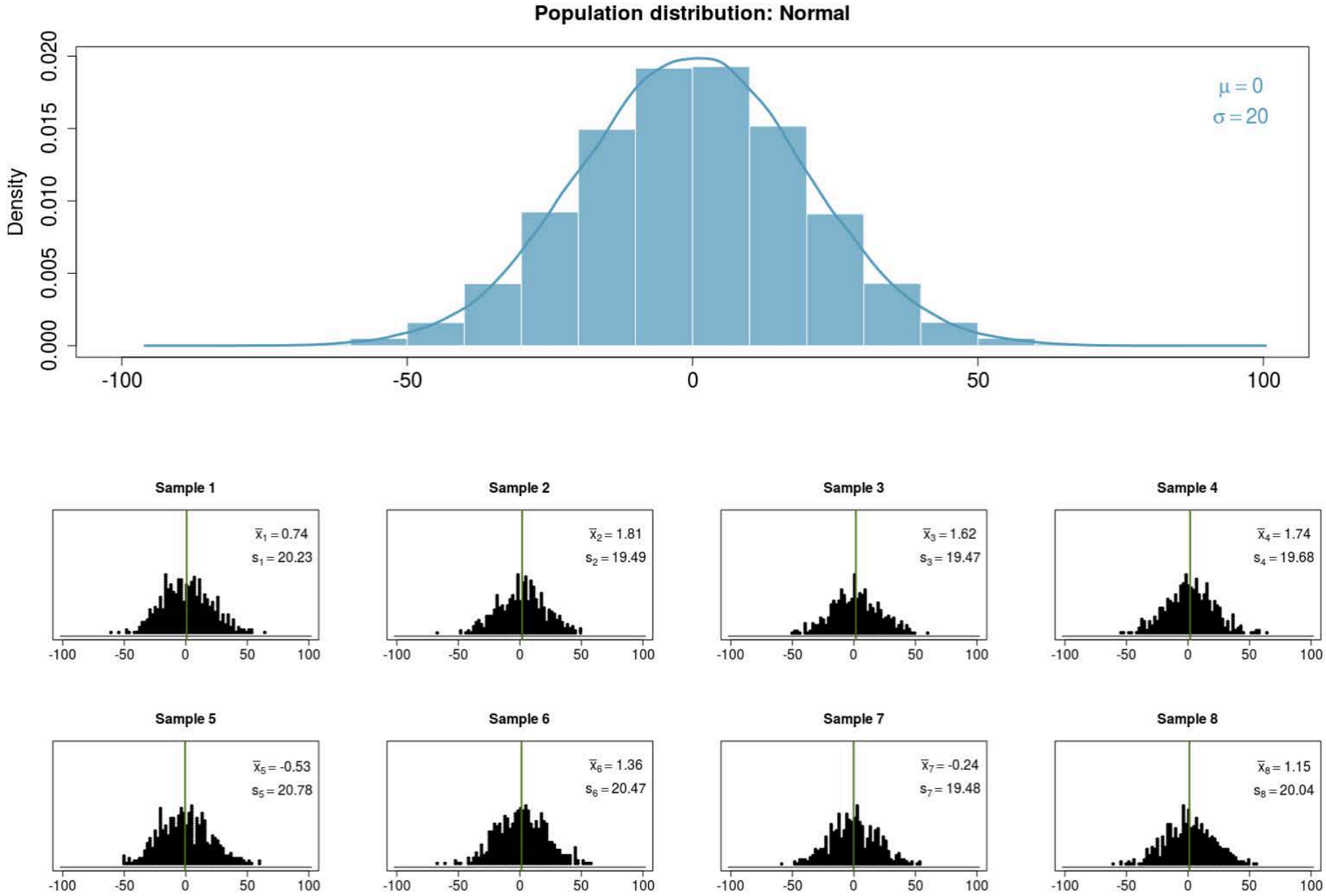


Distribution of means of 200 random samples, each consisting of 31 observations from a normal population

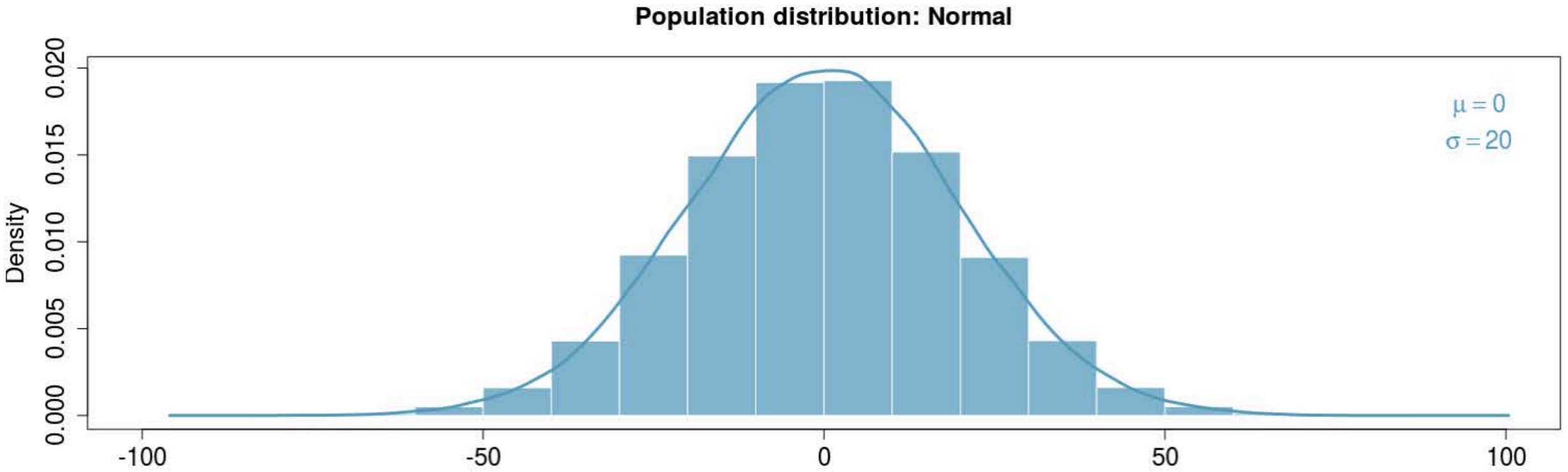
https://gallery.shinyapps.io/dist_calc/

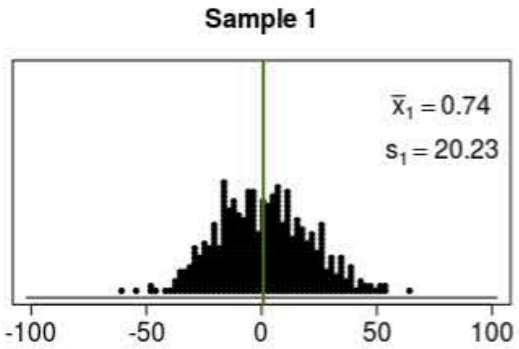


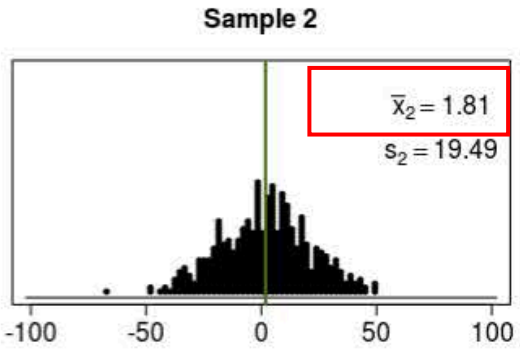
https://gallery.shinyapps.io/dist_calc/

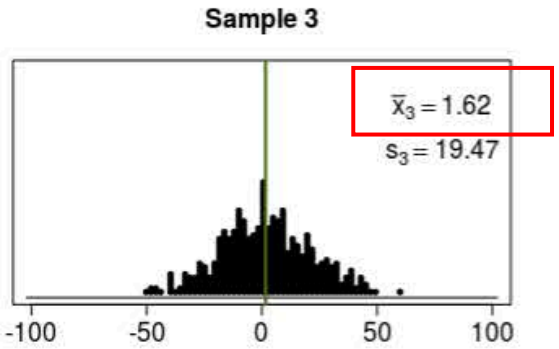


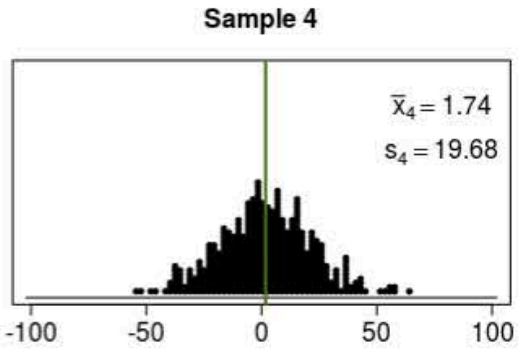
https://gallery.shinyapps.io/dist_calc/



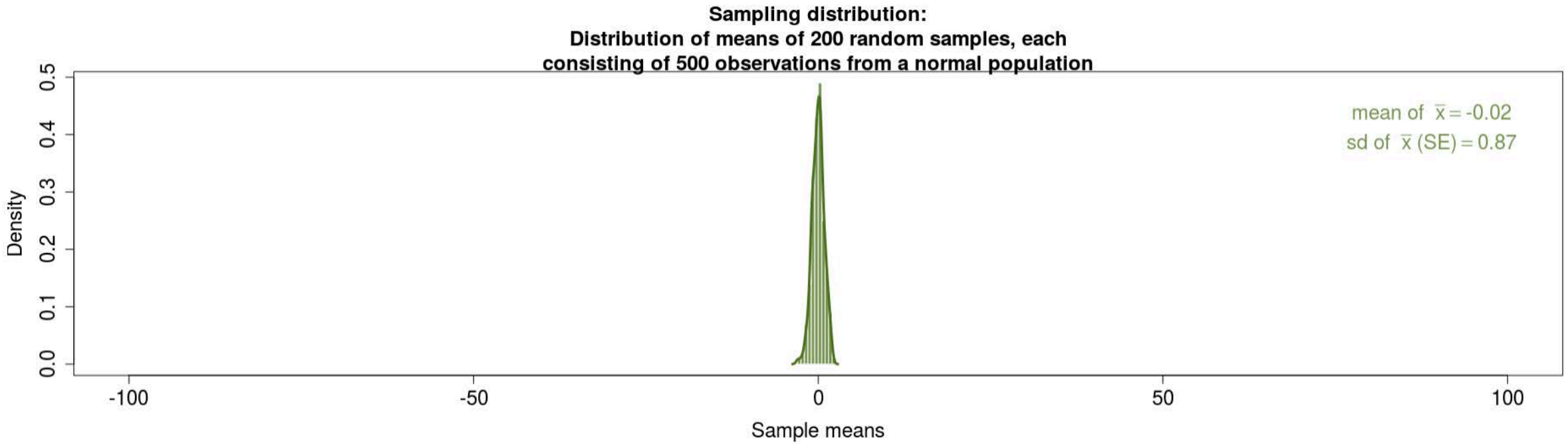








https://gallery.shinyapps.io/dist_calc/



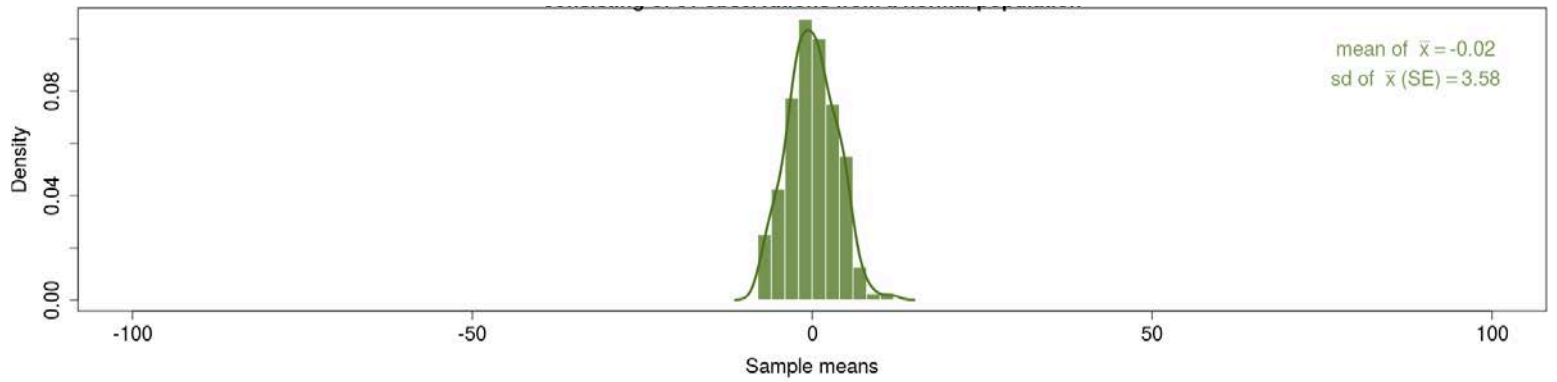
Distribution of means of 200 random samples, each consisting of 500 observations from a normal population

https://gallery.shinyapps.io/dist_calc/

Normally distributed samples

https://www.flickr.com/photos/jpallan/4633000725

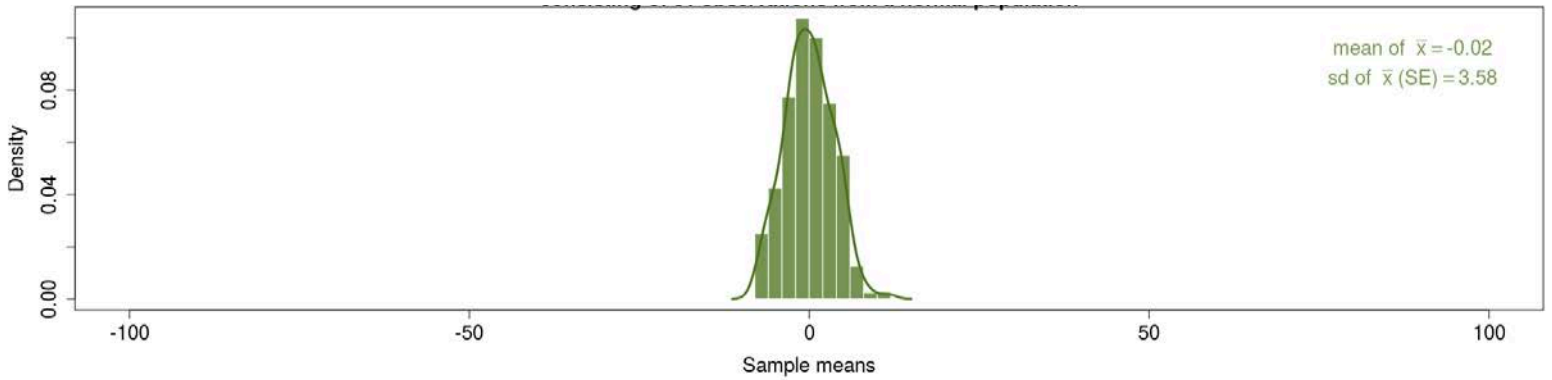
Central Limit Theorem (CLT)



$$\bar{x} \sim Normal\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

The distribution of sample statistics should be **normally distributed**, **centred** at the population mean (**μ**), and with a **standard deviation equal** to the **population standard deviation divided by the square root of the sample size**

Central Limit Theorem (CLT)



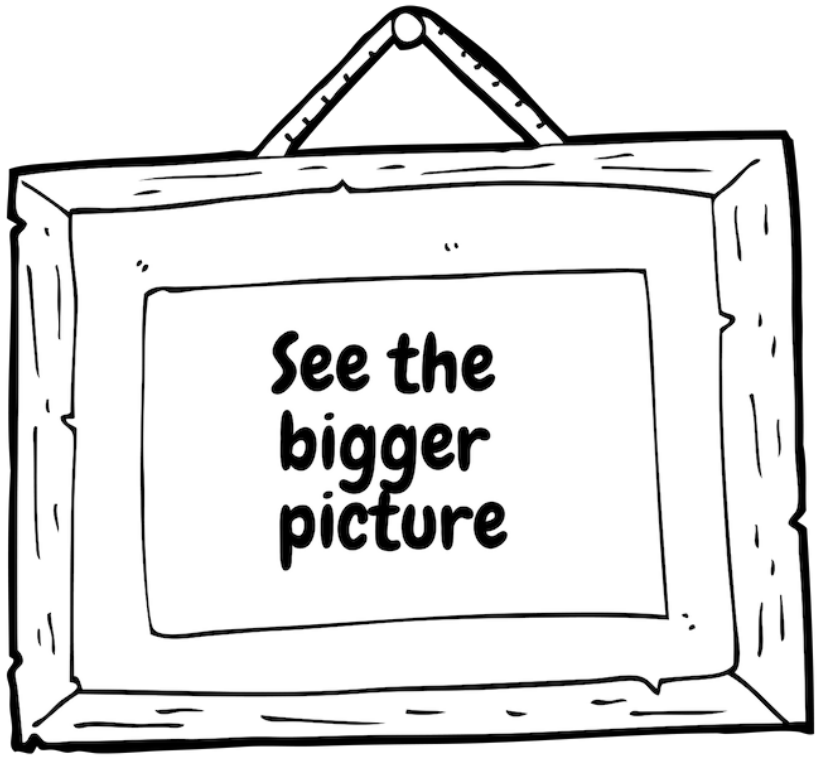
Conditions:

1. Observations must be independent (random sampling, n< 10% of population)

2. Either the population is normal, or if the population is skewed, the sample size must be large (*n>30)

Confidence Intervals (CI)

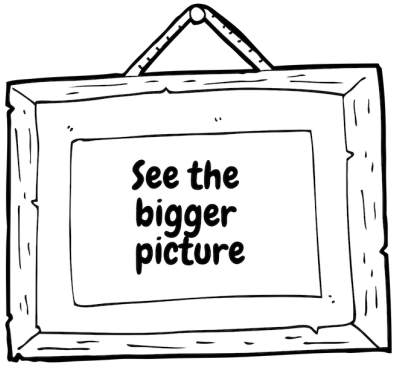
CI's are a **plausible range of values** of the population parameter



Confidence Intervals (CI)

Any interval will be constructed around the samples means ($\bar{x}$)

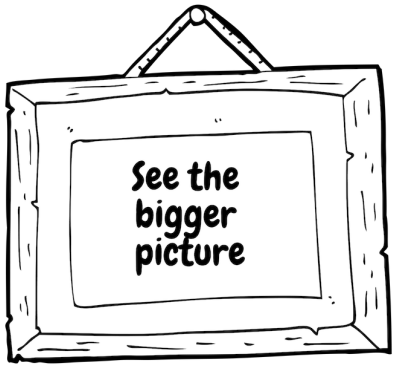
Which according to the CLT: are nearly normally distributed, and the centre of that distribution is at the unknown population mean



Confidence Intervals (CI)

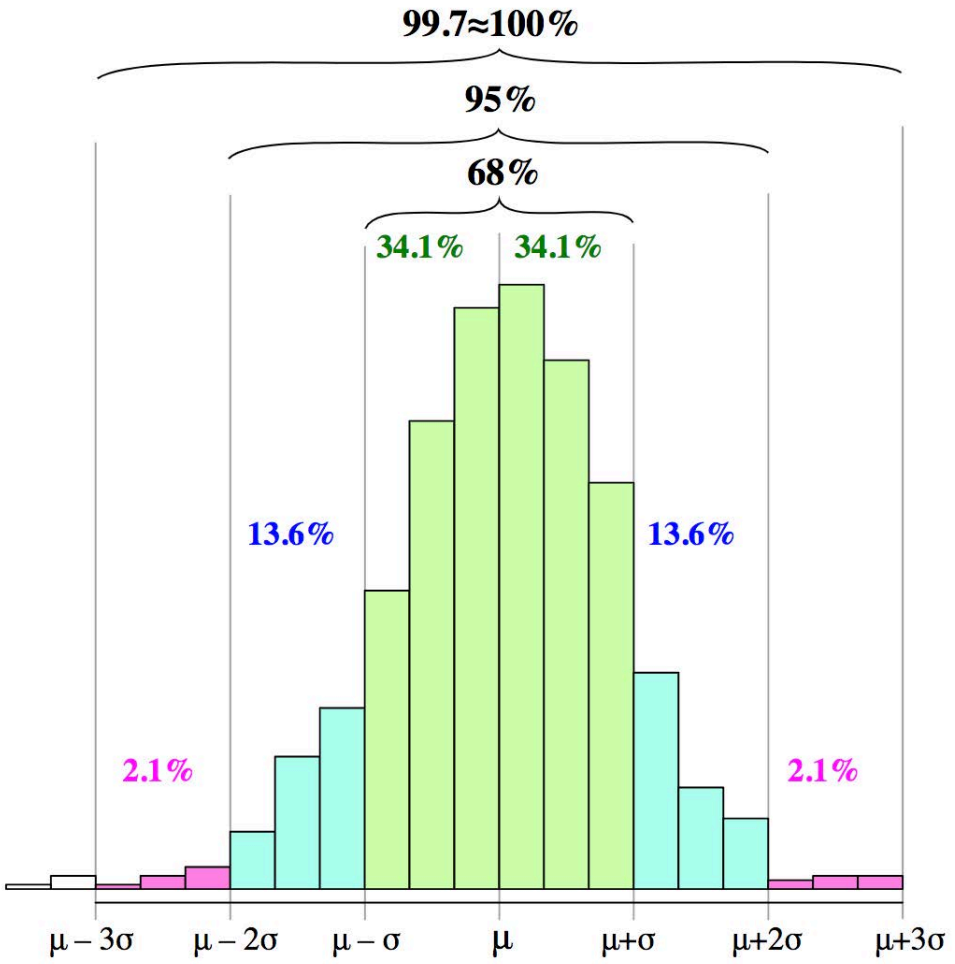
Any interval will be constructed around the samples means ($\bar{x}$)

Will tells us about the percentage of random samples that will yield confidence intervals that contain the true average number of the population



68, 95, 99 % empirical rule

This rule of thumb expresses that nearly all values taken lie within three standard deviations of the mean

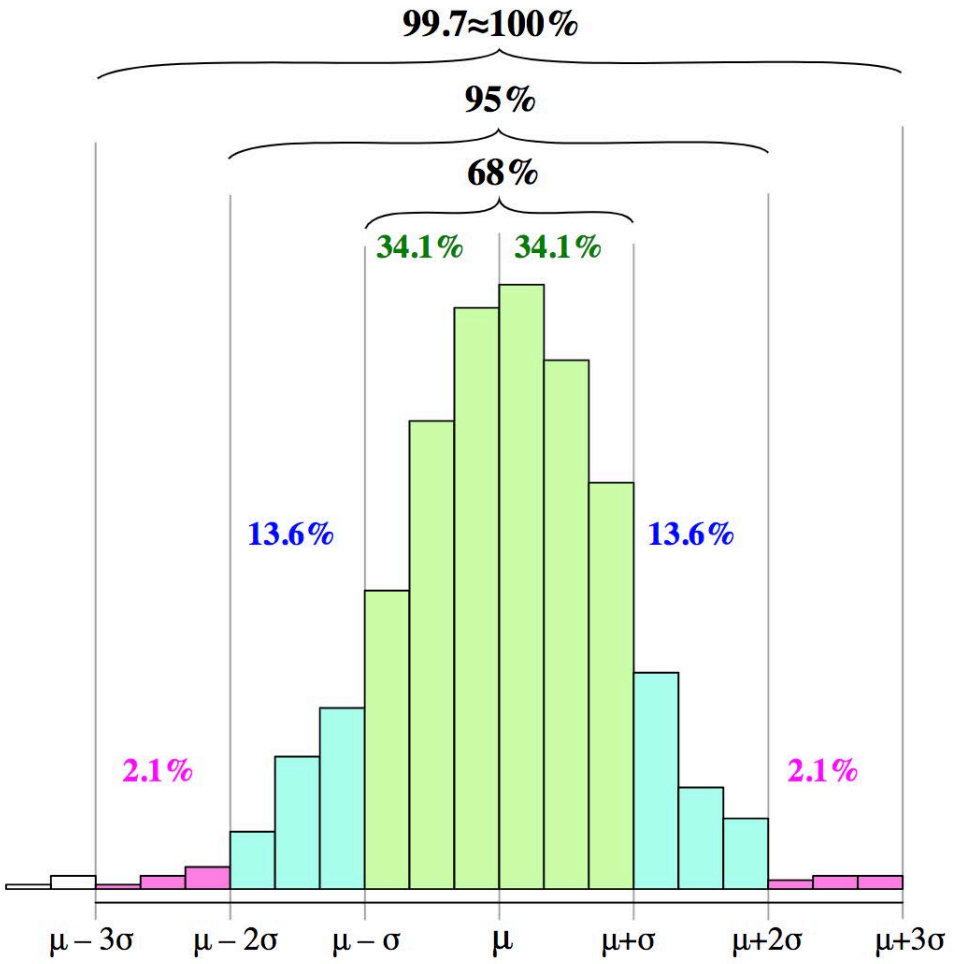


https://upload.wikimedia.org/wikipedia/commons/2/22/Empirical_rule_histogram.svg

68, 95, 99 % empirical rule

So for 95 % of random samples: The unknown true population mean is going to be within ≈ two standard errors of that sample's mean ($\bar{x} \pm 2\,SE$)

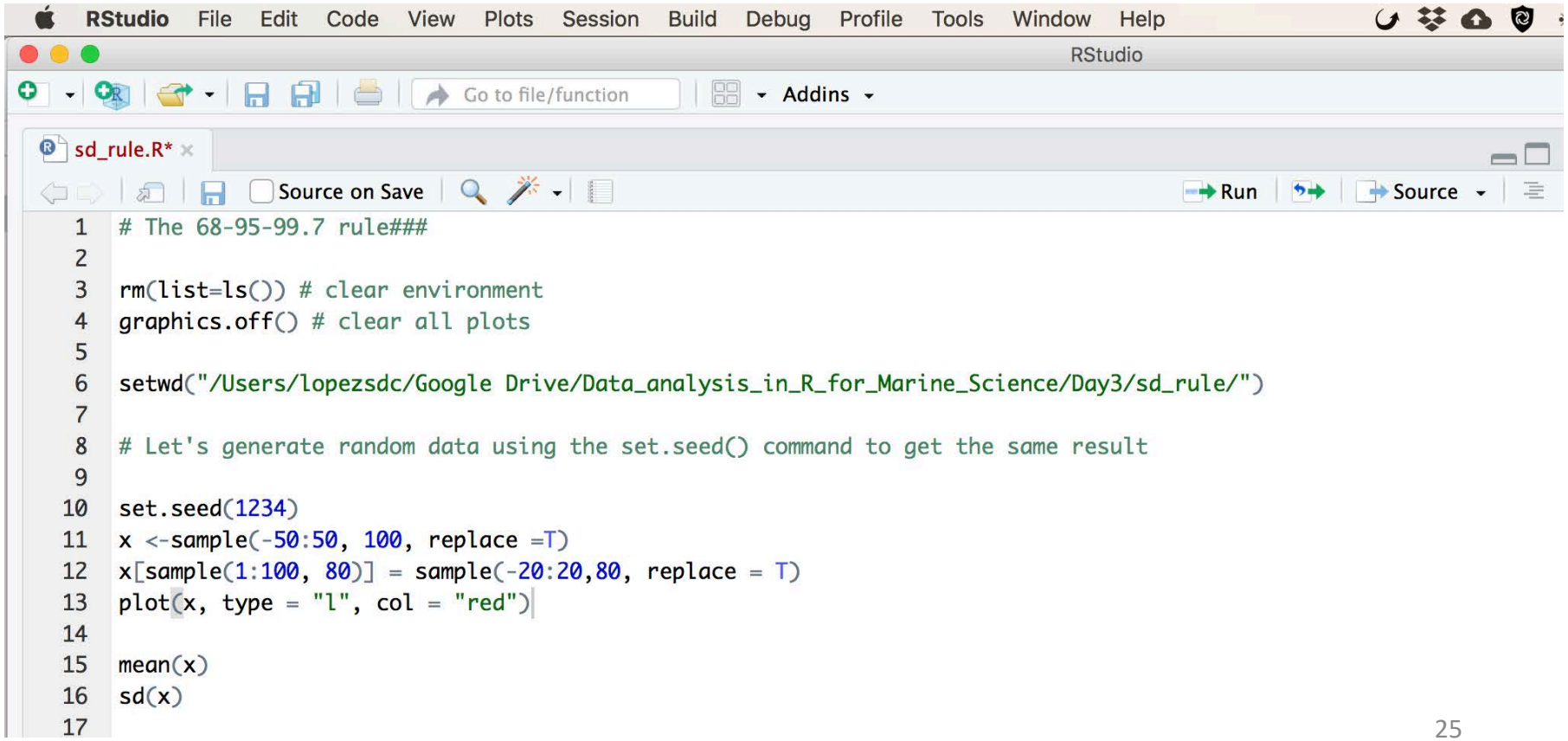
Margin of Error



https://upload.wikimedia.org/wikipedia/commons/2/22/Empirical_rule_histogram.svg

Understanding the empirical rule

1. You will have an example (sd_rule.R)



MENU

nature human behaviour

Comment | Published: 01 September 2017

Redefine statistical significance

Daniel J. Benjamin, James O. Berger, […] Valen E. Johnson

Nature Human Behaviour **2**, 6–10 (2018) | Download Citation

We propose to change the default P-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Remove, rather than redefine, statistical significance

Valentin Amrhein & Sander Greenland

Nature Human Behaviour **2**, 4 (2018) | Download Citation

Osteoarthritis and Cartilage

Volume 20, Issue 8, August 2012, Pages 805-808

Editorial

Why the *P*-value culture is bad and confidence intervals a better alternative

J. Ranstam ✉

Show more

https://doi.org/10.1016/j.joca.2012.04.001

Summary

In spite of frequent discussions of misuse and misunderstanding of probability values (*P*-values) they still appear in most scientific publications, and the disadvantages of erroneous and simplistic *P*-value interpretations grow with the number of scientific publications. Osteoarthritis and Cartilage prefer confidence intervals. This is a brief discussion of problems surrounding *P*-values and confidence intervals.

Confidence Intervals (CI)

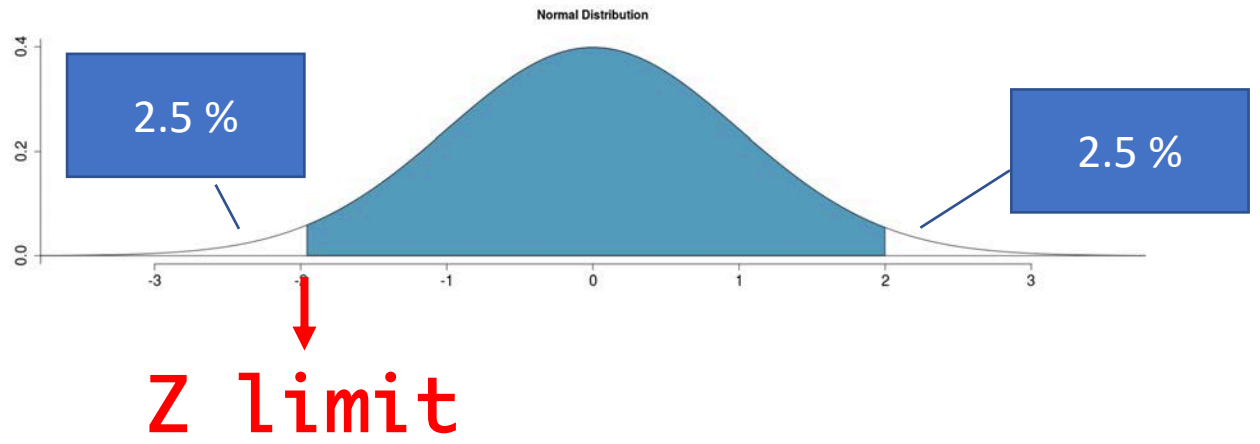
CI can be computed as the sample mean +/- a margin of error

$$\bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

The margin of error is value corresponding to the middle of the chosen percent (e.g., 95 %) of the normal distribution times the standard error of the sampling distribution

Confidence Intervals (CI)

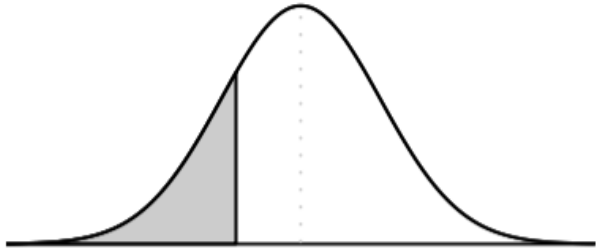
Finding the critical value 95% confidence using Z table



The area under curve below the lower bound of the middle 95 % is:

= 1-0.95 divided by 2 = 0.05/2 = **0.025**

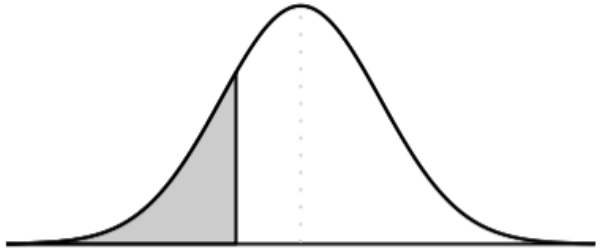
Normal probability table



Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9

The exact critical value for a 95 % CI is -1.96

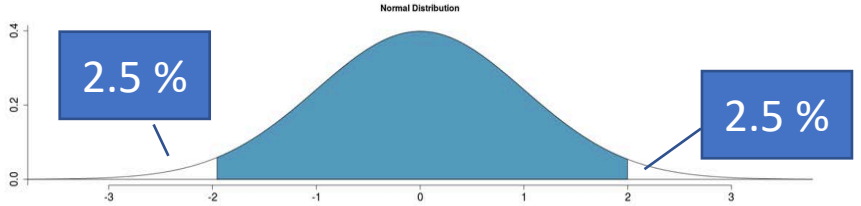
Normal probability table



Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9

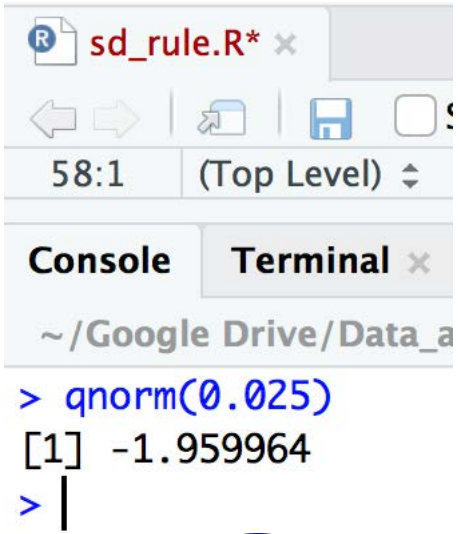
Finding the critical value 95% confidence

In R:



Use the function qnorm and the quantile (0.025) as an input

➢ `qnorm(0.025)`



The result is also negative, we just need to remember that we need the positive version of this number

Exercise

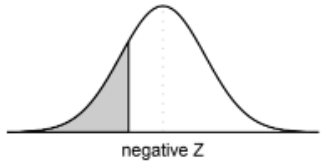
Use the qnorm function and find the critical value associated with a 90% confidence level

Or the Z table!

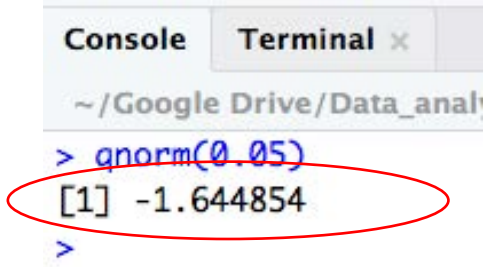
Exercise

1. Find the percentil (1 – 0.9/2 = 0.05)
2. Use the qnorm function = -1.64

Normal probability table



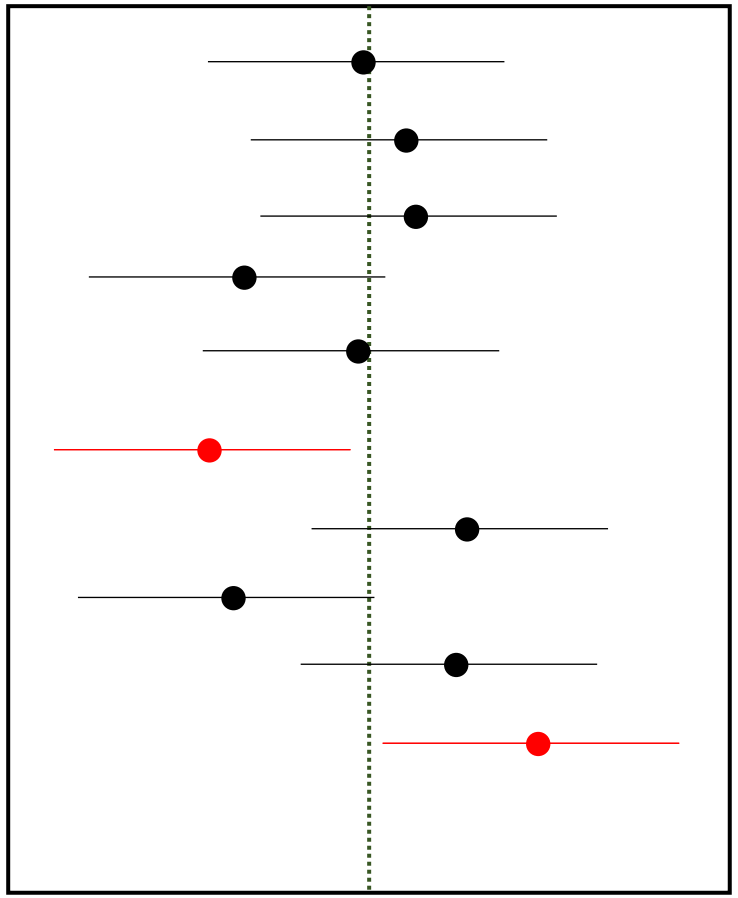
Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	1.5



One las thing about Confidence Intervals (CI)

Let's say we want to grasp the value of a population parameter such as its mean (μ).

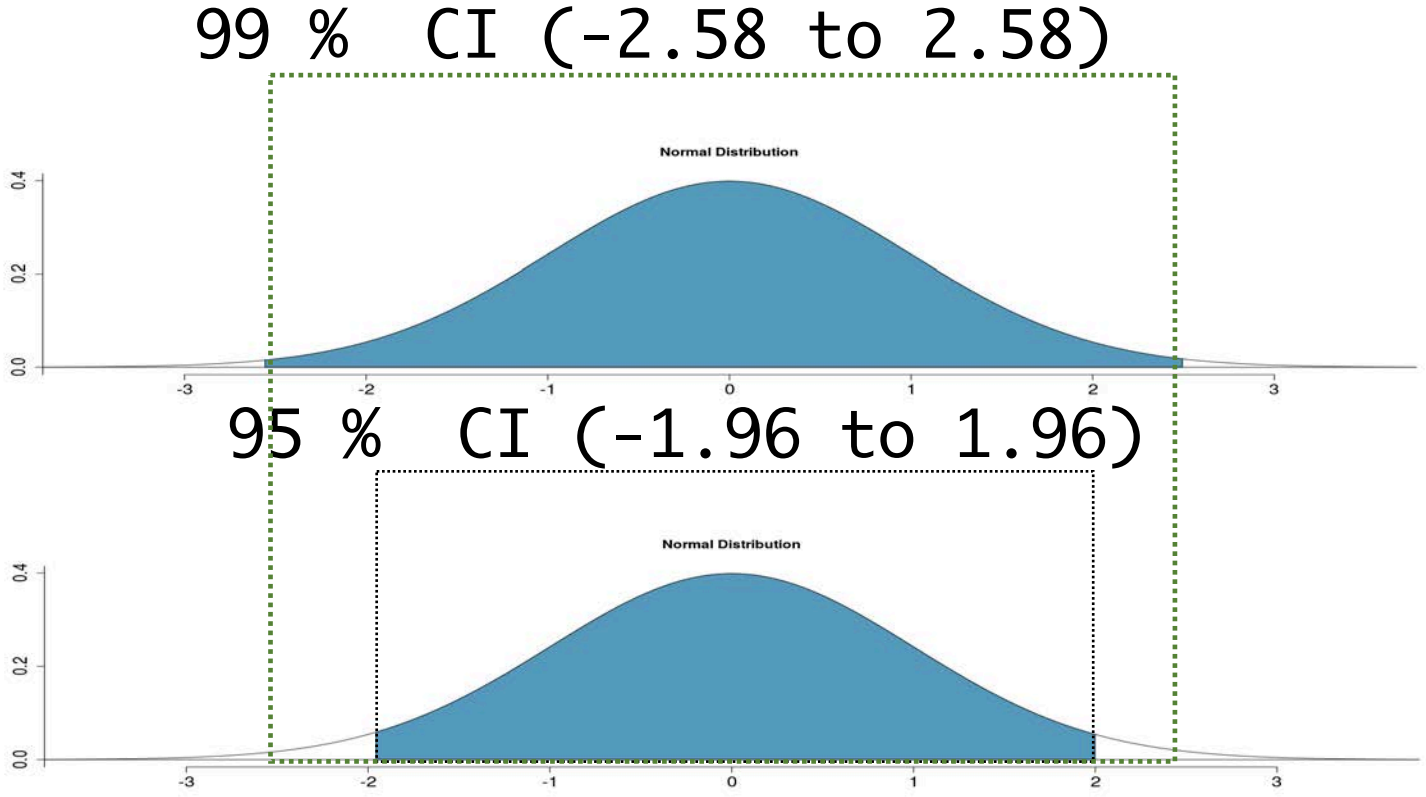
What shall we use, a narrow or a wide CI?



One las thing about Confidence Intervals (CI)

As the CI increases so does the width of the confidence interval

We can say then our measure is more accurate



Hypothesis Testing

https://commons.wikimedia.org/wiki/File:Tenniel_red_queen_with_alice.jpg

Hypothesis testing

A group of people at KAUST conducted a study to determine the amount of coffee that postdocs and PhD students from the campus drinks per day, and found that, on normal conditions (meaning that all people interviewed were well-rest and happy), the average daily consumption is 2 cups

Hypothesis testing

However, we might have observed that on regular basis postdocs and PhD students are not particularly relaxed, so we think the study is biased, and we actually believe that the average daily coffee consumption is higher than 2 cups

Our hypothesis is that $\mu \geq 2$

Hypothesis testing

When we state **our hypothesis**, we are **mainly proposing an explanation** for an observed phenomenon

However, there **might be another** possible **explanation** (also expressed as a hypothesis) **that invalidates** (annuls) **our** proposed **hypothesis**. This other idea will be stated as our **null hypothesis (Ho)**

Hypothesis testing

In our coffee example, **to invalidate** (annul) **our hypothesis** ($\mu \geq 2$) it suffices to say:

$$\textbf{Ho: } \mu = 2$$

We refer to our hypothesis (**the hypothesis we are investigating**) as the **alternative hypothesis** and is denoted as **Ha**

$$\textbf{Ha: } \mu \geq 2$$

The procedure for evaluating a hypothesis is called hypothesis testing

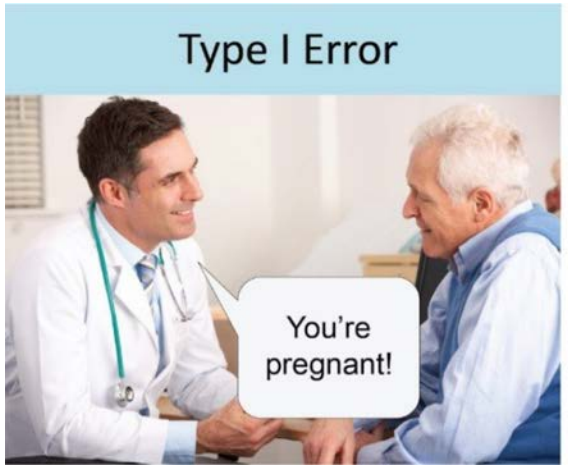
We might make two types of error

- Type I: We **reject Ho when it is true** and should not be rejected (it has a probability called α)

- Type II: **We fail to reject Ho when it is false** and should be rejected (its probability is called β)

Ho: You are not pregnant

Ha: You are pregnant



Which error is worse?

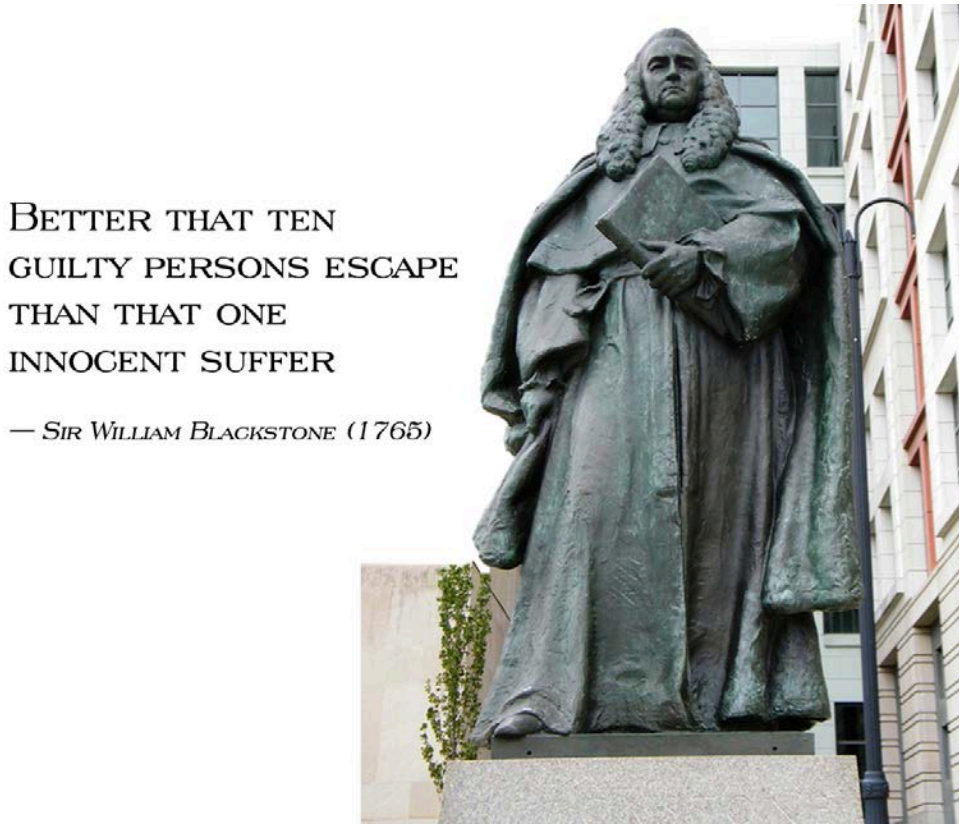
Ho: Innocent
Ha: Guilty

- A type two error where we make the defendant innocent when they're actually guilty?

- A type one error where we declare the defendant guilty when they are actually innocent?

https://www.cartoonstock.com/directory/d/decision_making_process.asp

Ho: Innocent
Ha: Guilty

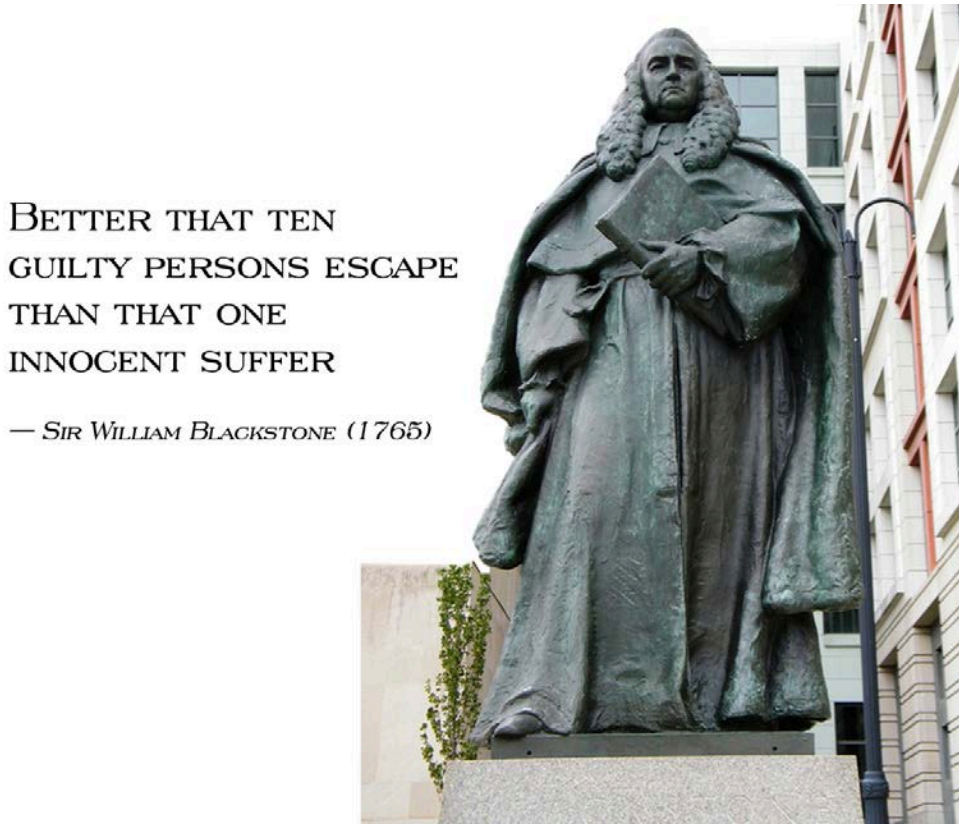


Kaz Vorpal

You have to make sure to minimize the rate of the type one error where we would be declaring a defendant guilty when is actually innocent.

https://ar.m.wikipedia.org/wiki/%D9%85%D9%84%D9%81:Blackstones-ratio.jpg

Ho: Innocent
Ha: Guilty



Kaz Vorpal

If you do not agree with this statement and you think that a type two error is a worst offense, then you would want to be minimizing the type two error rate

https://ar.m.wikipedia.org/wiki/%D9%85%D9%84%D9%81:Blackstones-ratio.jpg

Commonly, a type I error rate (alpha) of **0.01, 0.05 and 0.1 is tolerable**

Traditionally we reject the null hypothesis when the p-value is less than 0.05

Setting the significance level (our α) to 0.05

Means that for those cases where the null hypothesis is actually true, we do not want to incorrectly reject it more than 5%

In other words, when using a 5% significance level, there is about a 5% chance of making a type one error if the null hypothesis is true

- The **plausibility of null hypothesis** is **measured** with *a* **p-value** which is *a* **probability** that takes *a* value between 0 and 1

- The p-value also referred as the observed level of significance

- The p-value is directly proportional to the plausibility of the null hypothesis so:

The smaller the p-value the less plausible is the null hypothesis

- When a p-value larger than 10% is obtained, we conclude that there is no substantial evidence that the null hypothesis is not a plausible statement.

  Notice that we cannot conclude that the Ho has been proven

- If the null hypothesis is accepted then this simply means that the Ho is a plausible statement

Exercise

We interviewed 50 individuals (postdocs and PhD students) from the RSRC and determined that on average, the amount of coffee consumed is 2.5 cups per day. The standard deviation was 1.5

Exercise

1. Calculate the 95 % CI= $\bar{x} \pm z^* \frac{s}{\sqrt{n}}$
2. Based on the above information, determine if the data support our suspected hypothesis that postdocs and PhD students drinks on average more than 2 cups per day (μ>2)

Back to R: TIP Use qnorm function

We interviewed 50 individuals (postdocs and PhD students) from the RSRC and determined that on average, the amount of coffee consumed is 2.5 cups per day. The standard deviation was 1.5.

Remember:

We hypothesised that the postdocs and PhD students consume on average more than 2 cup
Our hypothesis is ($\mu > 2$)

To invalidate this hypothesis is enough to say ($\mu = 2$)

We interviewed 50 individuals (postdocs and PhD students) from the RSRC and determined that on average, the amount of coffee consumed is 2.5 cups per day. The standard deviation was 1.5.

1. State our hypotheses:

$$\text{Ho} = \mu = 2 \text{ vs. Ha} = \mu > 2$$

2. To calculate the 95 % CI = $\bar{x} \pm z^* \frac{s}{\sqrt{n}}$

   a) First obtain the percentile:
      z = 1 - 0.95/ 2 = 0.025

   b) Then by using the qnorm function: z=-1.96

We interviewed 50 individuals (postdocs and PhD students) from the RSRC and determined that on average, the amount of coffee consumed is 2.5 cups per day. The standard deviation was 1.5.

$$\text{CI} = \bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

95 % CI = 2.5 ± 1.96 (1.5/sqrt(50))

95 % CI = 2.5 ± 0.43

95 % CI = (2.1 to 2.9)

But what is the actual probability of rejection (p-value)?

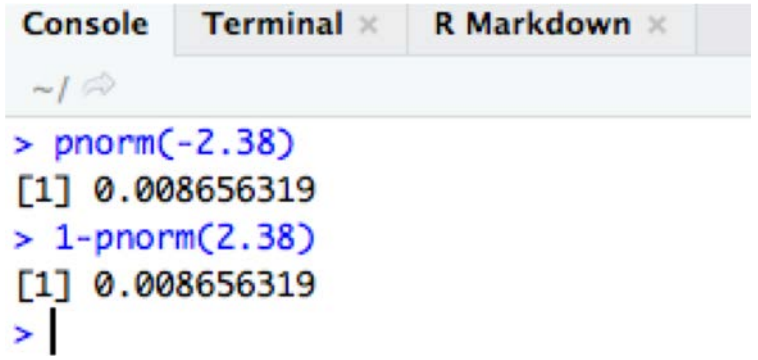
$$p(z) = \frac{\bar{x} - \mu}{SE} = \frac{2.5 - 2}{1.5/\sqrt{50}} = 2.38$$

Now we go back to the z table or use the pnorm function in R

- pnorn(-2.38)

Or

> 1-pnorm(2.38)



But what is the actual probability of rejection (p-value)?

$$p(z) = \frac{\bar{x} - \mu}{SE} = \frac{2.5 - 2}{1.5/\sqrt{50}} = 2.38$$

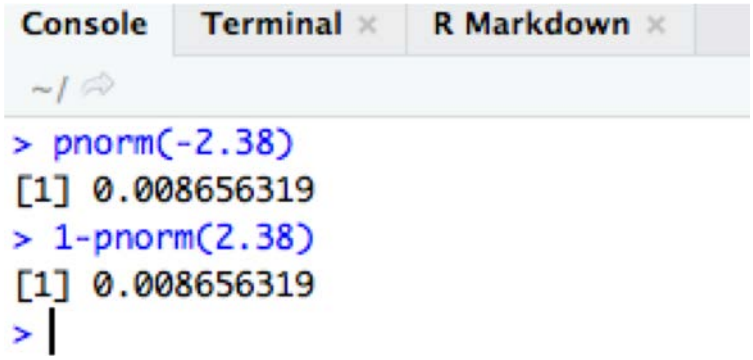
Now we go back to the z table or use the pnorm function in R

- `pnorn(-2.38)`

Or

`> 1-pnorm(2.38)`

p-value = 0.008



Our probability is (p-value) 0.008

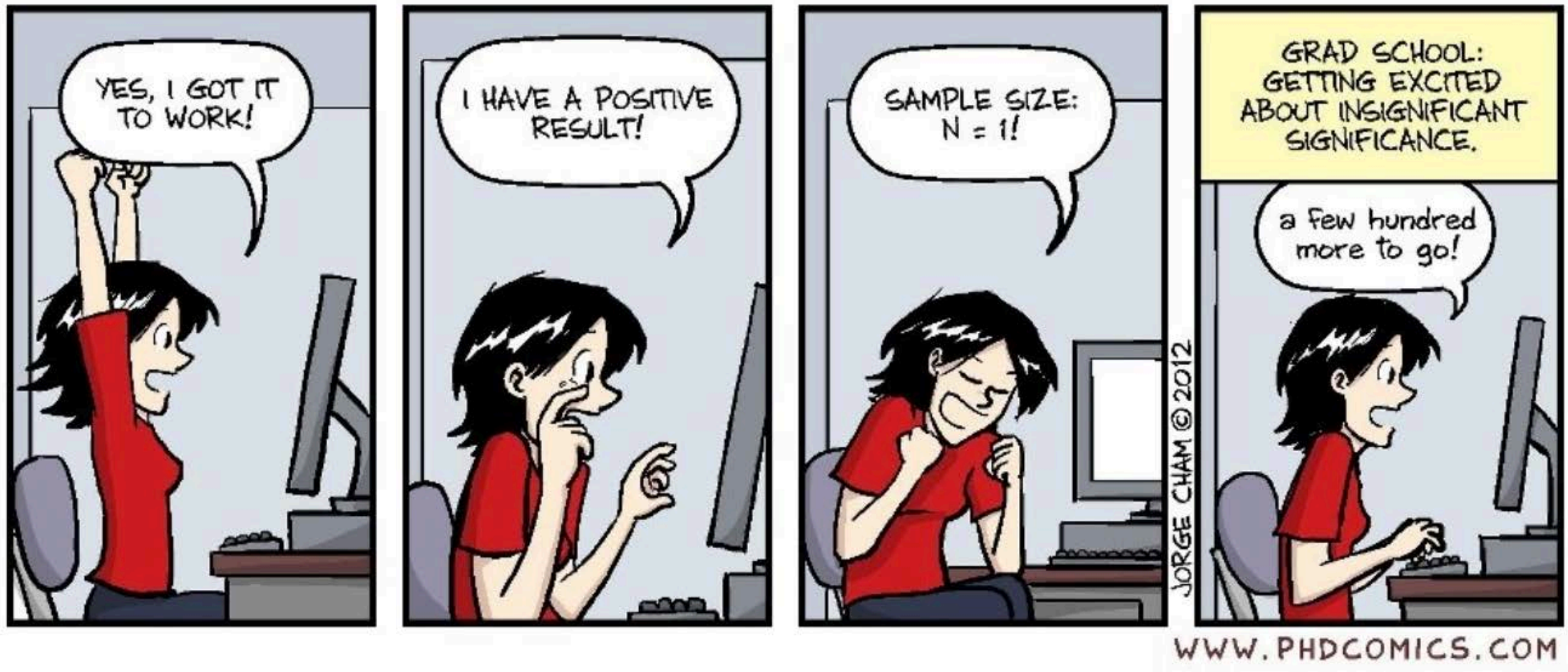
Which is lower than the significance level (alpha 0.05).

So we have enough evidence to reject Ho.

Meaning that it will be **unlikely to see our data if the Ho was true**

There is <1 % chance that from a random sample of 50 researchers (postdocs and students) will yield a mean of 2

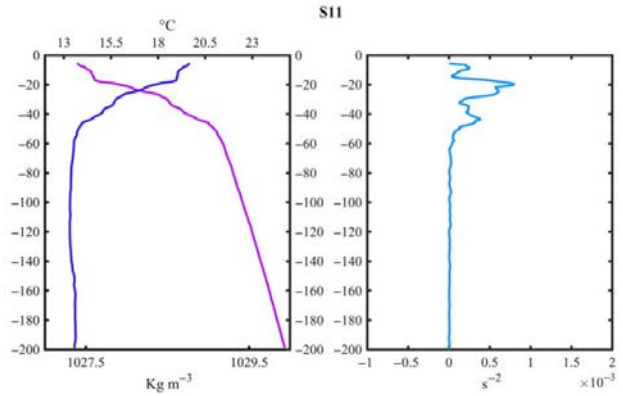
Variables

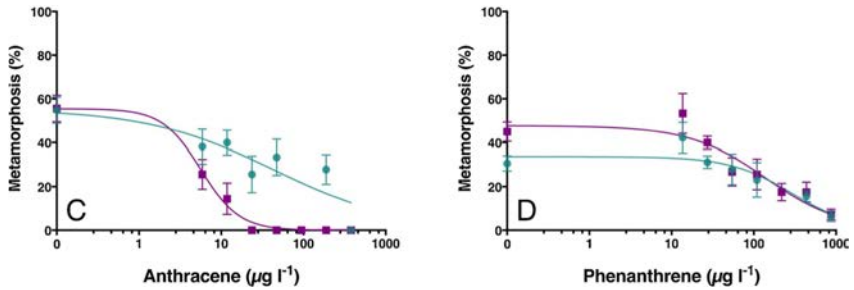

YES, I GOT IT TO WORK!
I HAVE A POSITIVE RESULT!
SAMPLE SIZE: N = 1!
GRAD SCHOOL: GETTING EXCITED ABOUT INSIGNIFICANT SIGNIFICANCE.
a few hundred more to go!
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

Variables

Any characteristics, numbers or quantities that can be measured or counted

- Numerical → Quantitative
  - Continuous →
  - Discrete →





Images from: Lopez-Sandoval et al., 2018 and *Overmans et al., 2018*

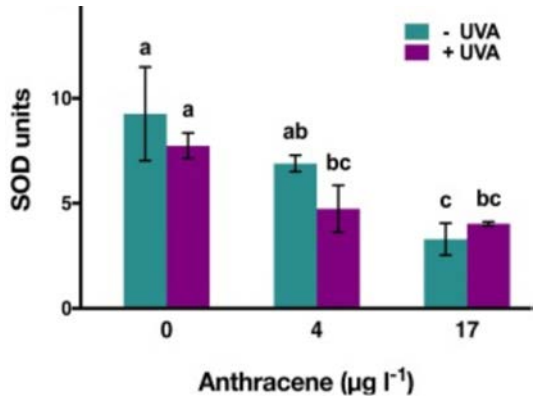
Variables

Any characteristics, numbers or quantities that can be measured or counted

- Categorical
  - Ordinal
  - Nominal

Bivariate Distributions for Ordinal Variables (Frequencies)

	VOTING				COMPLEX				NOCARE			
NOSAY	AS	A	D	DS	AS	A	D	DS	AS	A	D	DS
AS	70	51	28	14	84	52	17	10	81	60	19	3
A	82	313	88	9	123	303	57	9	106	292	93	1
D	93	280	413	25	90	496	202	23	44	286	457	24
DS	30	14	51	28	17	54	33	19	10	18	67	28



Images from: Jöreskog 2012, and *Overmans et al., 2018*

Data display for R

LONG FORMAT 

y=abundance	x=treatment1	replicate
10^5	PLUS	1
10^5	PLUS	2
10^5	PLUS	3
10^4	INSITU	1
10^4	INSITU	2
10^4	INSITU	3
10^3	MINUS	1
10^3	MINUS	2
10^3	MINUS	3

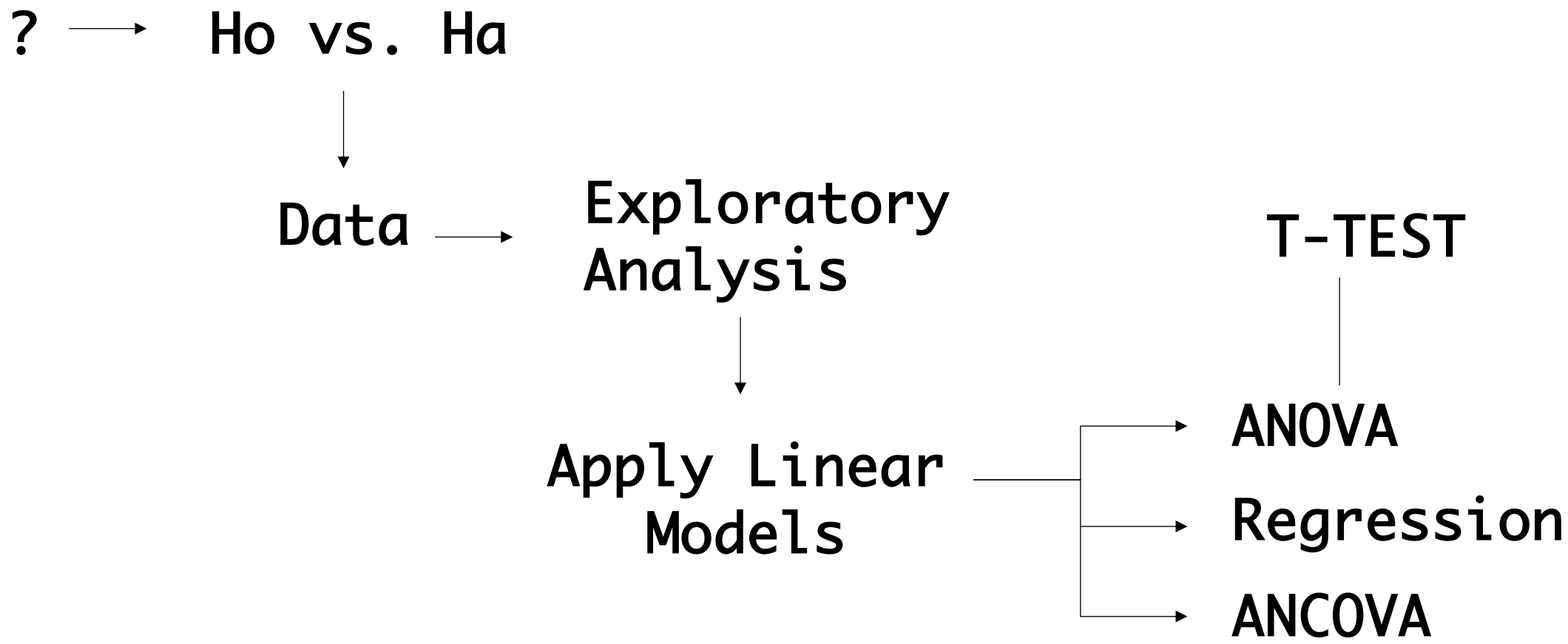
WIDE FORMAT 

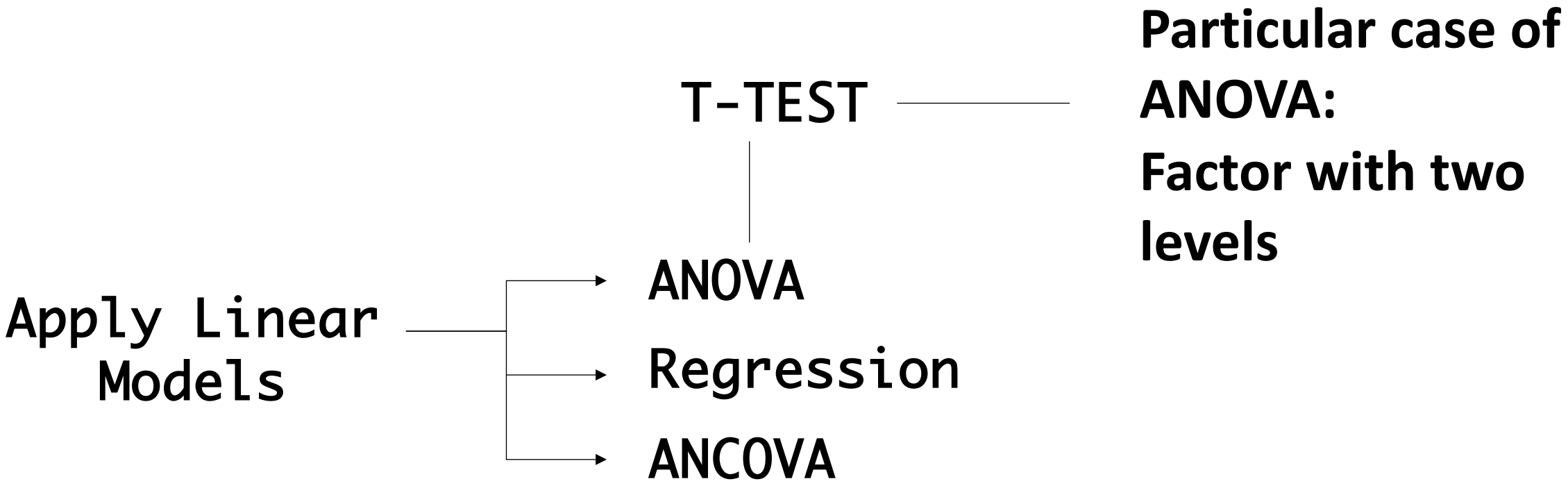
x=treatment1	replicate1	replicate2	replicate3
PLUS	10^5	10^5	10^5
INSITU	10^4	10^4	10^4
MINUS	10^3	10^3	10^3

Save format as CSV
(Excell also works, but requires more steps)
No weird characters
No spaces (use . _ - instead)

Each row represents one individual observation

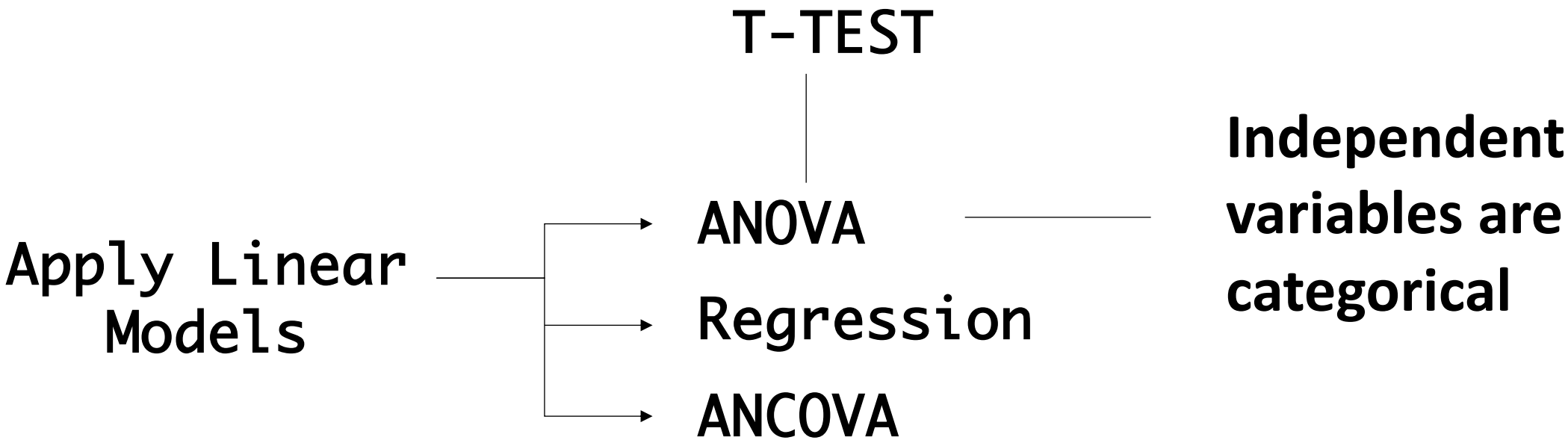
Linear Models


?
Ho vs. Ha
Data
Exploratory Analysis
Apply Linear Models
T-TEST
ANOVA
Regression
ANCOVA



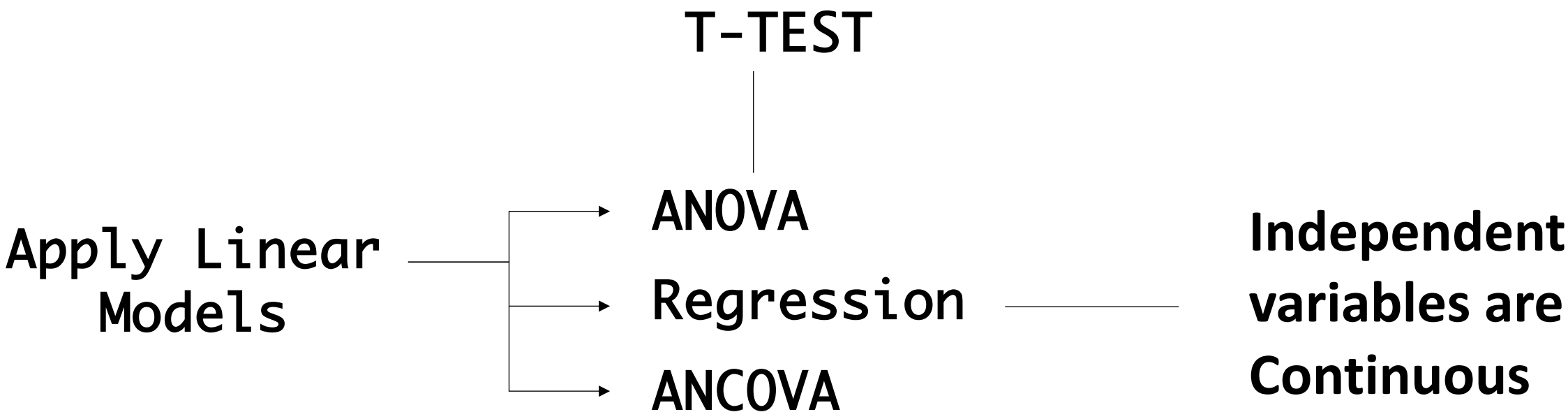
Dependent (Response): Growth rates

Independent (Explanatory): Temperature (High, Low)



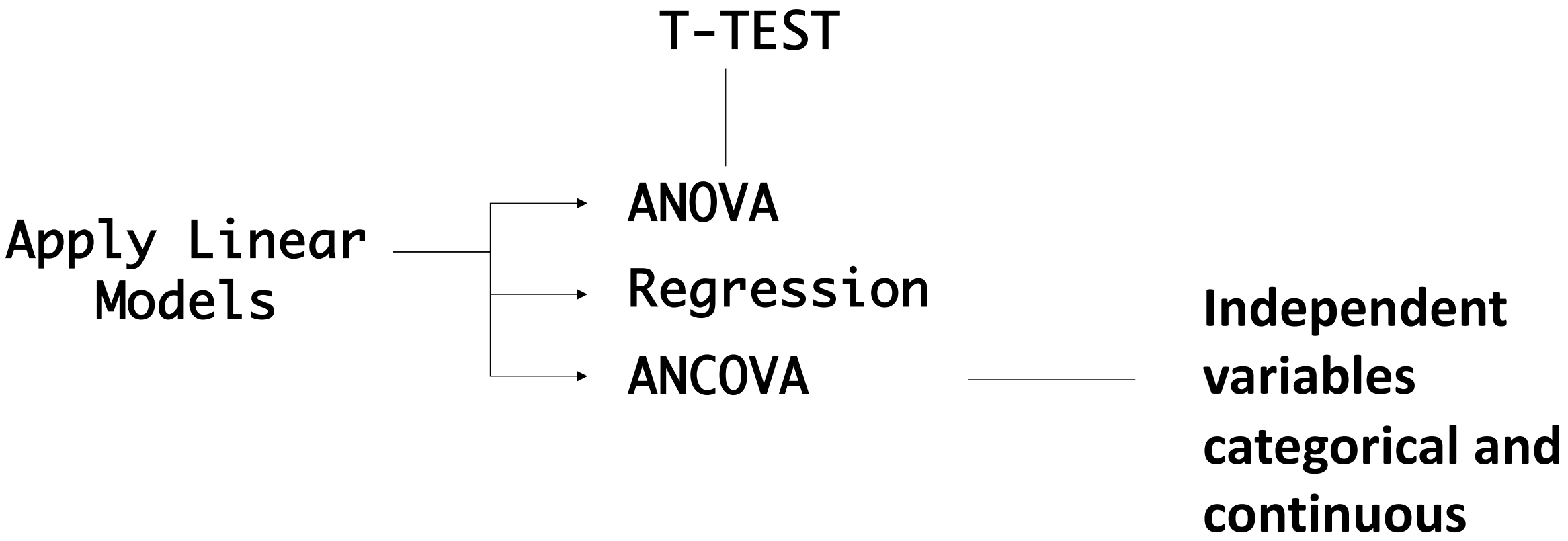
Dependent (Response): Growth rates (several experiments)

Independent (Explanatory): Temperature (High, Low)



Dependent (Response): Growth rates

Independent (Explanatory): Nutrient concentration



Dependent (Response): Growth rates

Independent (Explanatory): Nutrient concentration (cont) and Temp (High, low)

T-test: comparing two groups

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Statistical Inference for the relationship between two variables

T distribution

- To describe the distribution of the sample mean when the population standard deviation is unknown
- **To compare two means when the population standard deviations are unknown**

The standard deviation of our sampling distribution =

$$\cdot\bar{x}_{1 \text{ and } 2} \rightarrow \text{SD}_{1,2}=\sqrt{{}^{s1}/_{n1} + {}^{s2}/_{n2}} \rightarrow SE_{1,2}$$

Conditions

I. Independence

- Independence within groups
  - Random sample
  - If sampling without replacement, n < 10% of population
- Independence between groups
  - Meaning that the two groups must be independent of each other (non-paired)

Conditions

II. Increase sample size if skew

- The more skew the population distributions, the higher the sample size needed

Similar conditions to what we saw CLT, why?

Because we **assume** samples come from a **normally distributed population**

NORMALITY TESTS

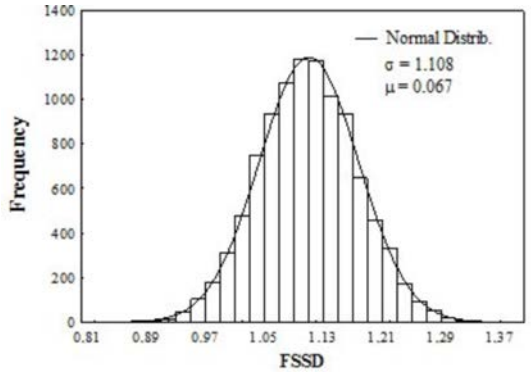
VISUALIZATION

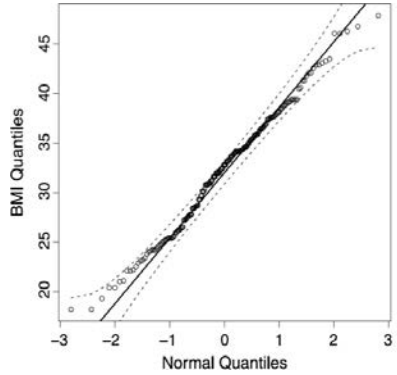
TEST

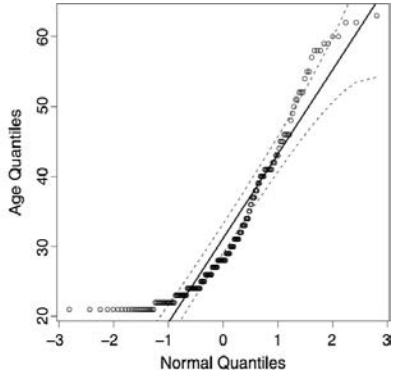
HISTOGRAM

Q-Q PLOTS

Shapiro-Wilk



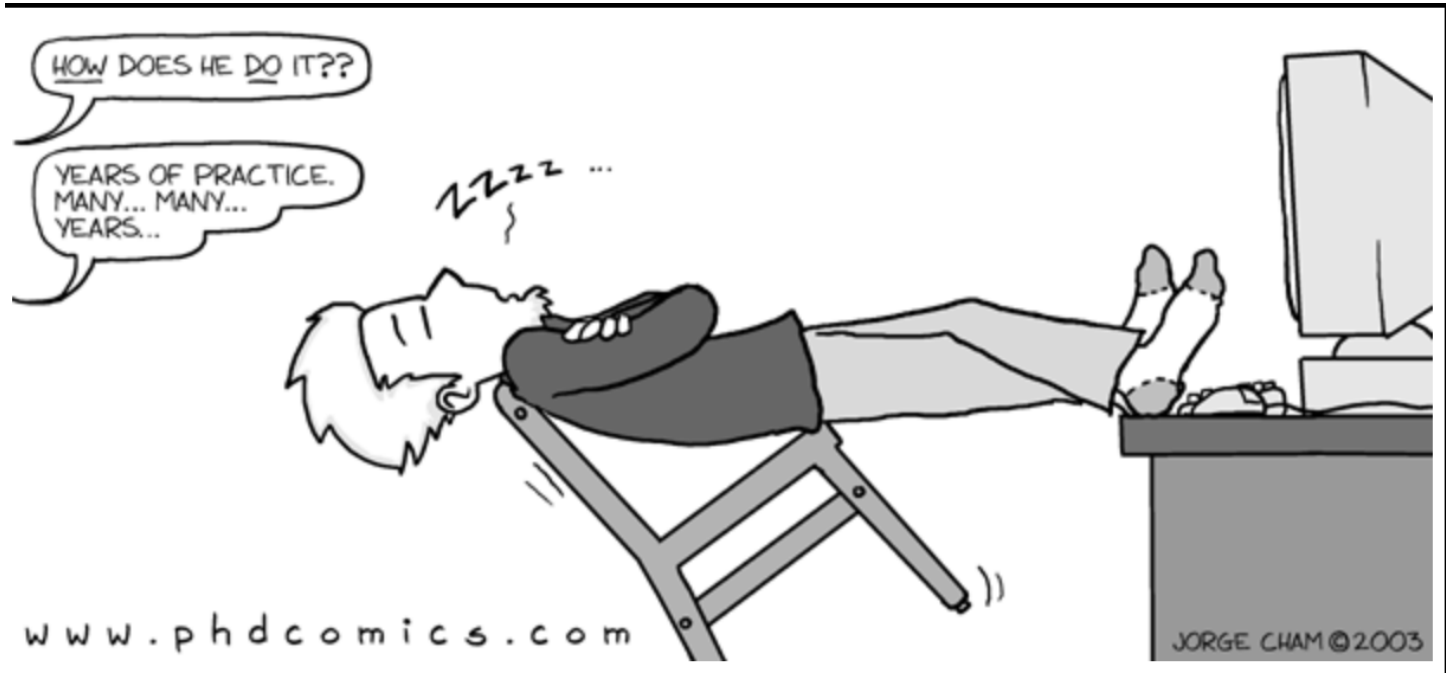




Ho : $\mu = normal$

Ha: $\mu \neq normal$

R Practice
Comparing two groups



Two-Sample t-test for comparing the means

1. Download the file t_testdata.csv to your working directory for day 3

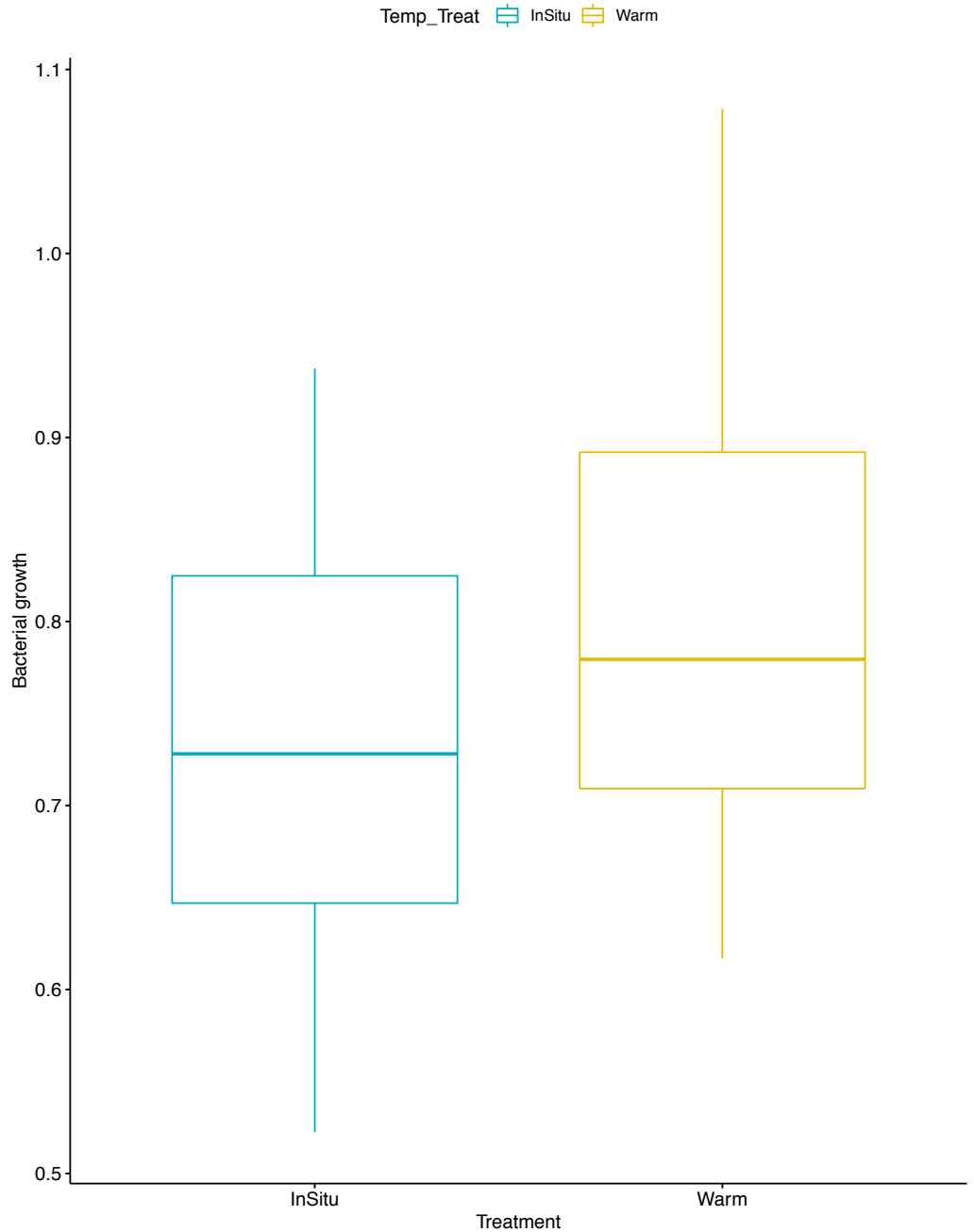
2. Open 

Two-Sample t-test

Determine if bacterial growth rates where affected by the temperature treatment (InSitu and Warm)

Our hypotheses:

Ho, $\mu_1 - \mu_2 = 0$
Ha, $\mu_1 - \mu_2 \neq 0$


Temp_Treat
InSitu
Warm
1.1
1.0
0.9
0.8
0.7
0.6
0.5
Bacterial growth
InSitu
Warm
Treatment

Two-Sample t-test

Function in R

1. Subset our sample

```
grInSituW <- gr[gr$Temp_Treat %in% c("InSitu","Warm"),]
```

2. T-TEST

```
t.test(grInSituW$GrowthRate ~ grInSituW$Temp_Treat)
```

```
        Welch Two Sample t-test

data:  grCW$GrowthRate by grCW$Temp_Treat
t = -1.9047, df = 29.082, p-value = 0.06675
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.178427138  0.006337807
sample estimates:
mean in group InSitu   mean in group Warm
           0.7243086             0.8103533
```

Result is **not statistically significant**
Therefore, fail to reject Ho (mu1-mu2=0)
and conclude that significant difference does not exist in growth rates within temperature treatments

```
        Welch Two Sample t-test

data:  grCW$GrowthRate by grCW$Temp_Treat
t = -1.9047, df = 29.082, p-value = 0.06675
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.178427138  0.006337807
sample estimates:
mean in group InSitu   mean in group Warm
           0.7243086             0.8103533
```

At 0.95 confidence level, we believe that the true difference between the two means falls between −0.17 and 0.006

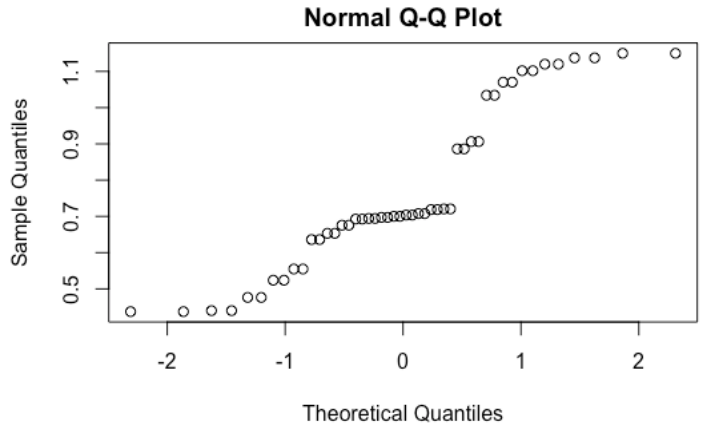
Remember a p-value

Does not mean that **there is a 95% chance** that a given hypothesis **is correct**

It **signifies** that **if the null hypothesis is true**, and all other assumptions are valid, **there is a 5 % chance of obtaining a result at least as extreme as the one observed**

What if our data does not follow a normal distribution?

Non-parametric test (Wilcoxon-Mann-Whitney)



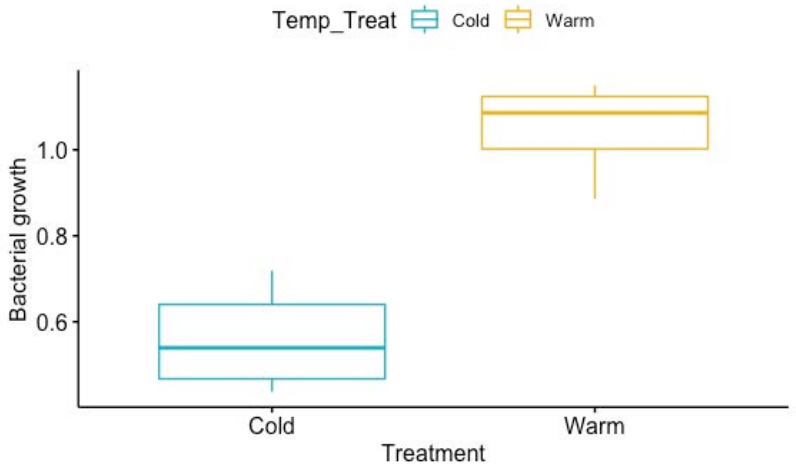
Shapiro-Wilk normality test data: gr$GrowthRate_NP W = 0.89245, p-value = 0.0003621

Non-parametric test (Wilcoxon-Mann-Whitney)

Null hypothesis is that the means of both treatments are equal

```
grCW <- gr[gr$Temp_Treat %in% c("Cold","Warm"),]

wilcox.test(GrowthRate_NP ~ Temp_Treat, data = grCW,
            exact = FALSE)

ggboxplot(grCW, x = "Temp_Treat", y = "GrowthRate_NP",
          color = "Temp_Treat", palette = c("#00AFBB", "#E7B800"),
          ylab = "Bacterial growth", xlab = "Treatment")
```

Non-parametric test (Wilcoxon-Mann-Whitney)



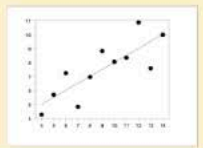
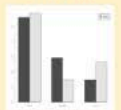
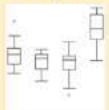
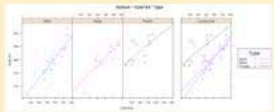
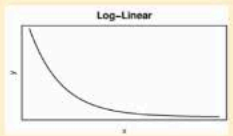
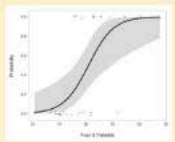
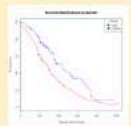

```
Wilcoxon rank sum test with continuity correction data:
GrowthRate_NP by Temp_Treat W = 0, p-value = 1.491e-06 alternative
hypothesis: true location shift is not equal to 0
```

Comparing more than two groups

https://www.highspeedinternet.com/resources/the-most-streamed-shows/

Independent (x)

Explanatory Variable

Dependent Variable (y) / Response Variable	Continuous	Categorical	Both
Continuous Real numbers	**REGRESSION** Scatter plot 	**ANOVA** Bar chart  Box plot 	**ANCOVA** Coplot 
Count Integer >=0	**Log-LINEAR MODEL** 	**CONTINGENCY TABLE** Event observed / Event forecast: Yes, No, Marginal total Yes: a, b, a+b No: c, d, c+d Marginal total: a+c, b+d, a+b+c+d=n	**ANALYSIS OF DEVIANCE** Analysis of Maximum Likelihood Estimates Parameter / DF / Estimate / Standard Error / Wald Chi-Square / Pr > ChiSq Intercept 1 -1.3922 0.1724 65.2041 <.0001 B scout 1 0.0289 0.3793 0.0058 0.9392 S medium 1 -0.4948 0.2955 2.8048 0.0940 S high 1 -1.9921 0.7394 7.2597 0.0071 B*S scout medium 1 -0.1472 0.5315 0.0767 0.7818 B*S scout high 1 0.1568 0.8893 0.0311 0.8601
Proportion data 0>x>1	**LOGISTIC REGRESSION** 	**CONTINGENCY TABLE** Event observed / Event forecast: Yes, No, Marginal total Yes: a, b, a+b No: c, d, c+d Marginal total: a+c, b+d, a+b+c+d=n	**ANALYSIS OF DEVIANCE** Analysis of Maximum Likelihood Estimates Parameter / DF / Estimate / Standard Error / Wald Chi-Square / Pr > ChiSq Intercept 1 -1.3922 0.1724 65.2041 <.0001 B scout 1 0.0289 0.3793 0.0058 0.9392 S medium 1 -0.4948 0.2955 2.8048 0.0940 S high 1 -1.9921 0.7394 7.2597 0.0071 B*S scout medium 1 -0.1472 0.5315 0.0767 0.7818 B*S scout high 1 0.1568 0.8893 0.0311 0.8601
Age of death Survival analysis	**GLM** 	**ANALYSIS OF DEVIANCE** Analysis of Maximum Likelihood Estimates Parameter / DF / Estimate / Standard Error / Wald Chi-Square / Pr > ChiSq Intercept 1 -1.3922 0.1724 65.2041 <.0001 B scout 1 0.0289 0.3793 0.0058 0.9392 S medium 1 -0.4948 0.2955 2.8048 0.0940 S high 1 -1.9921 0.7394 7.2597 0.0071 B*S scout medium 1 -0.1472 0.5315 0.0767 0.7818 B*S scout high 1 0.1568 0.8893 0.0311 0.8601	**ANALYSIS OF DEVIANCE** Analysis of Maximum Likelihood Estimates Parameter / DF / Estimate / Standard Error / Wald Chi-Square / Pr > ChiSq Intercept 1 -1.3922 0.1724 65.2041 <.0001 B scout 1 0.0289 0.3793 0.0058 0.9392 S medium 1 -0.4948 0.2955 2.8048 0.0940 S high 1 -1.9921 0.7394 7.2597 0.0071 B*S scout medium 1 -0.1472 0.5315 0.0767 0.7818 B*S scout high 1 0.1568 0.8893 0.0311 0.8601

Modified from A. Anton

Analysis of Variance models (ANOVA)

Extension of the t-test used to:

Compare the means of multiple groups identified by a categorical variable with more than two possible categories

Analysis of Variance models (ANOVA)

Here our variables:

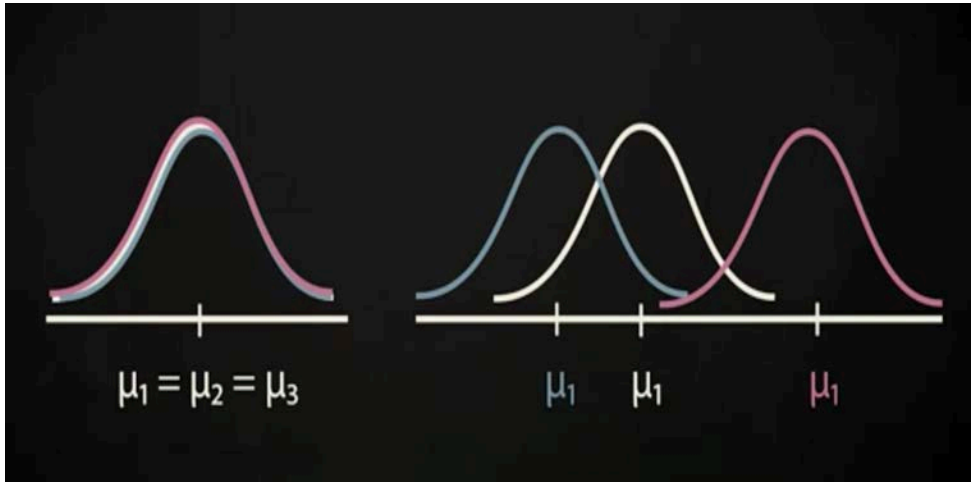
- Categorical variable is called **the factor** and is typically considered as the explanatory variable
- The **numerical variable**, whose means across different groups are compared, **is** regarded as **the response variable**

Example: Do chlorophyll values change in different months?
  x: month (12 levels)
  y: chlorophyll concentration

Example: Does bacterial growth rate change with temperature?
  x: temperature treatment (3 levels: PLUS, MINUS, IN SITU)
  y: bacterial growth rate (u-1)

Working hypothesis

- Ho: There are no differences between the means of the different groups
- Ha: At least one pair of means are different from each other



Ho: mu1=mu2=mu3

Ha: at least one pair of means are different form each other

Conditions for ANOVA

I. Independence

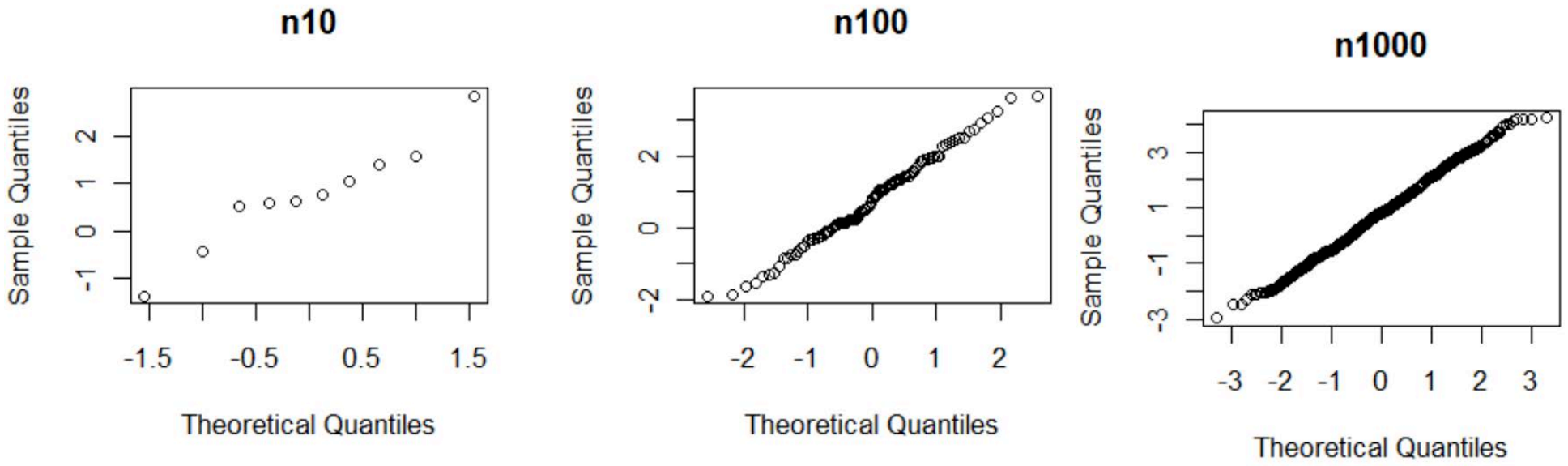
- Independence within groups: sampled observations must be independent

- Independence between groups
  - Meaning that the two groups must be independent of each other (non- paired)

Conditions for ANOVA

II. Approximate normality

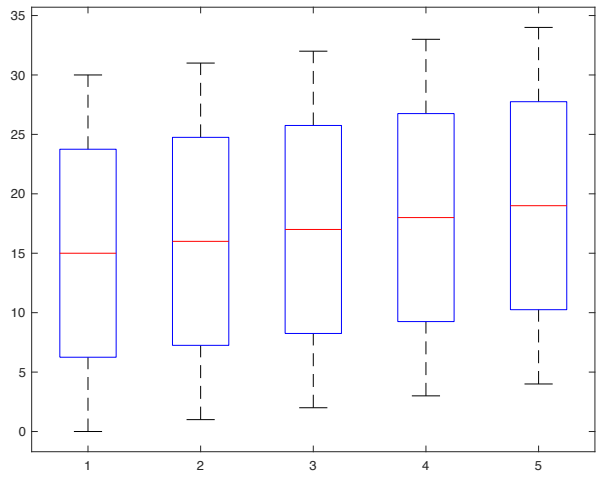
- Distributions should be nearly normal within each group, specially when sample sizes are small



Conditions for ANOVA

III. Constant Variance

Variability should be consistent across groups: **homocedastic** groups



Group	n	Mean	Std Dev
1	7	15	10.8
2	7	16	10.8
3	7	17	10.8
4	7	18	10.8
5	7	19	10.8
Overall	35	17	10.8

Levene's test
Df = 4,30
p =1

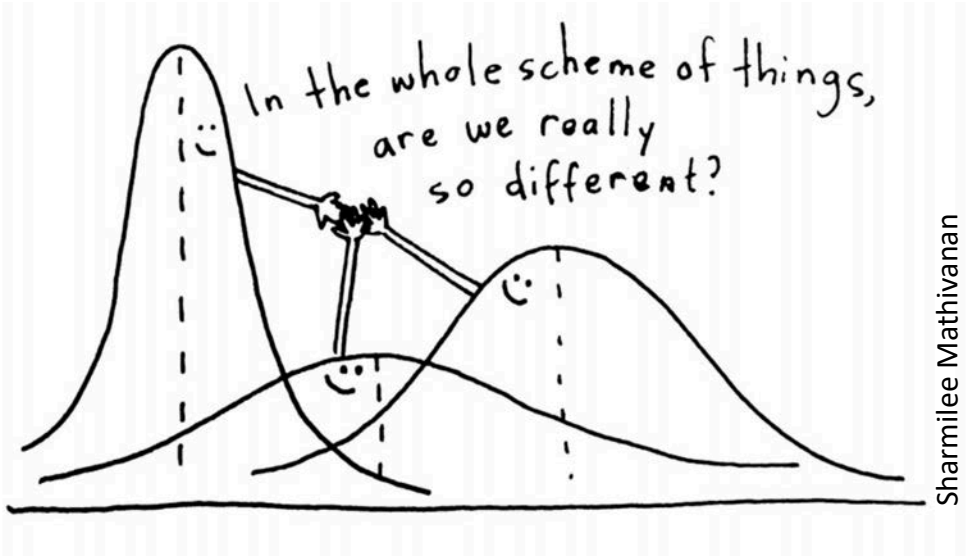
The test for examining the Ho is called ANOVA F-statistics and is defined:

$$F = \frac{Variability\ between\ groups}{Variability\ within\ in\ groups}$$

In order to reject Ho, we need a small p-value, which requires a large F statistics

Obtaining a large F statistics requires that the variability between sample means is greater than the variability within the samples

ANOVA with one factor (one-way ANOVA)



http://the-statistics-of-life.blogspot.com/2017/01/one-way-anova-welch-test.html

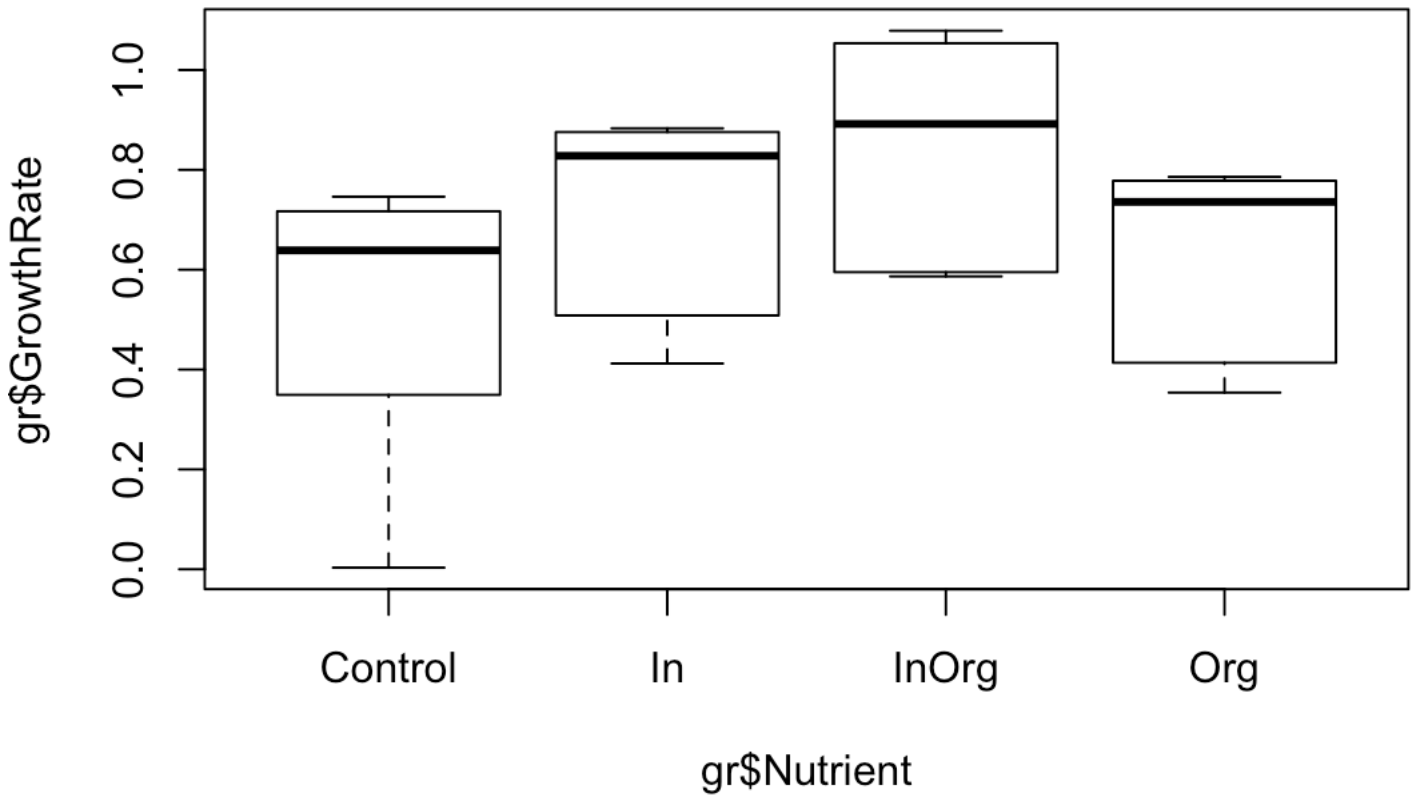
One-way ANOVA with R

1. Download the file ANOVA_data.csv to your working directory for day 3

2. Open 

1. Explore your data

```
plot(gr$GrowthRate ~ gr$Nutrient)
```



```
plot(gr$GrowthRate ~ gr$Nutrient)

anova2 <- aov(gr$GrowthRate ~ gr$Nutrient)

summary(anova2)

TukeyHSD(anova2)

plot(TukeyHSD(anova2))
```

```
> anova2 <- aov(gr$GrowthRate ~ gr$Nutrient)
> summary(anova2)
             Df Sum Sq Mean Sq F value Pr(>F)
gr$Nutrient   3 0.3585 0.11952   2.258  0.113
Residuals    20 1.0588 0.05294
```

Summ of squares of groups: Measures the variability between groups

The variability explained: square deviation of group means from overall mean, weighted by sample size

```
> anova2 <- aov(gr$GrowthRate ~ gr$Nutrient)
> summary(anova2)
             Df Sum Sq Mean Sq F value Pr(>F)
gr$Nutrient   3 0.3585 0.11952   2.258  0.113
Residuals    20 1.0588 0.05294
```

Total 1.4173

Summ of squares of error: Measures the variability within groups

Unexplained variability: by the group or due to other reasons. SSE= 1.4173-0.3585 = 1.0588

```
> anova2 <- aov(gr$GrowthRate ~ gr$Nutrient)
> summary(anova2)
             Df Sum Sq Mean Sq F value Pr(>F)
gr$Nutrient   3 0.3585 0.11952   2.258  0.113
Residuals    20 1.0588 0.05294
```

Mean squares: Average variability between and within groups, calculated as the total variability (SSQ) scaled to the associated Df

Group (MSG) SSG/dfG = 0.3585/3 = 0.1195
Error (MSE) SSE/dfE = 1.0588/20 = 0.0529

```
> anova2 <- aov(gr$GrowthRate ~ gr$Nutrient)
> summary(anova2)
             Df Sum Sq Mean Sq F value Pr(>F)
gr$Nutrient   3 0.3585 0.11952   2.258  0.113
Residuals    20 1.0588 0.05294
```

F statistics: Ratio of the average between group and within group variabilities:

F= MSG/MSE = 0.11952/0.0529 =2.258

```
> anova2 <- aov(gr$GrowthRate ~ gr$Nutrient)
> summary(anova2)
             Df Sum Sq Mean Sq F value Pr(>F)
gr$Nutrient   3 0.3585 0.11952   2.258  0.113
Residuals    20 1.0588 0.05294
```

You can also verify this value with the **pf function in R**

- `pf(2.258,3,20,lower.tail=FALSE)`
  `[1] 0.112915`

```
> anova2 <- aov(gr$GrowthRate ~ gr$Nutrient)
> summary(anova2)
             Df Sum Sq Mean Sq F value Pr(>F)
gr$Nutrient   3 0.3585 0.11952   2.258  0.113
Residuals    20 1.0588 0.05294
```

Since our p>0.05 we say that we have week evidence against the Ho, so we fail to reject the null hypothesis, and we conclude that Ho (mu1=mu2=mu3) is plausible

What if our data does not follow a normal distribution?

ANOVA with two factors (two-way ANOVA)

Two-way ANOVA (ANOVA with two factors)

In many two-way ANOVA procedures, one of the two factors is the main explanatory variable of interest

E.g., In our study of growth rates, we might believe temperature is the main driver controlling bacterial growth

Two-way ANOVA (ANOVA with two factors)

However, nutrients may also play a role there…

therefore we need to explore the effect of the interaction

Foods and condiments

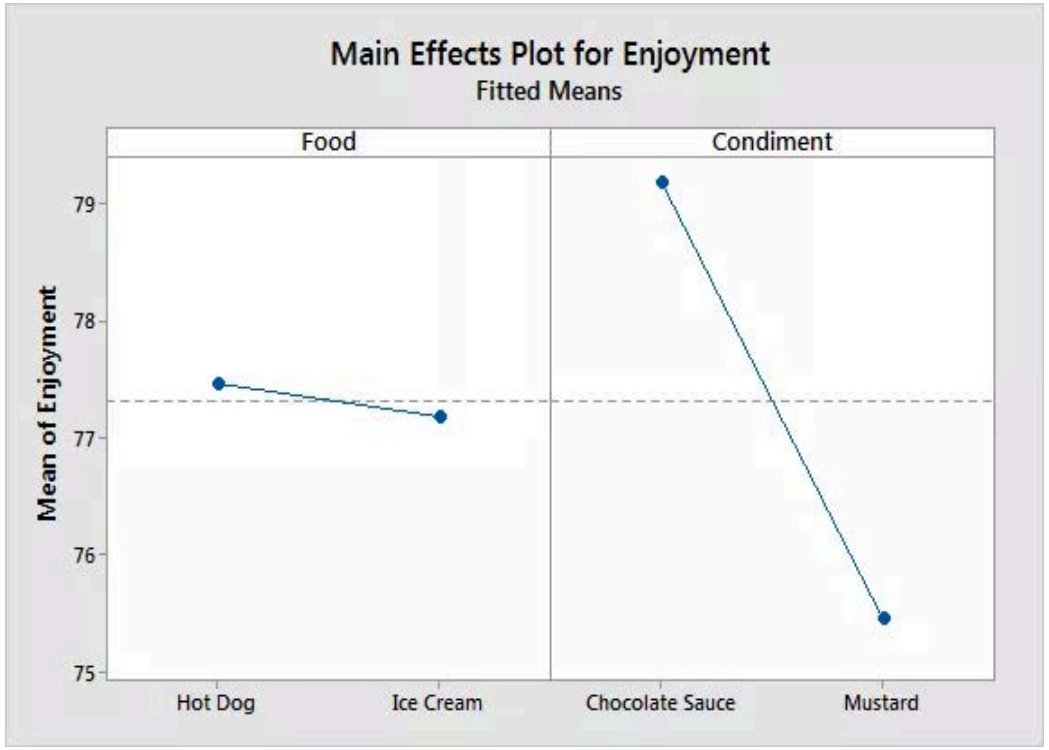
Ho = People like hotdogs and Ice cream the same

Ho = People like chocolate sauce and mustard the same

Results:

Food = People like Hot Dogs the same as Ice Cream

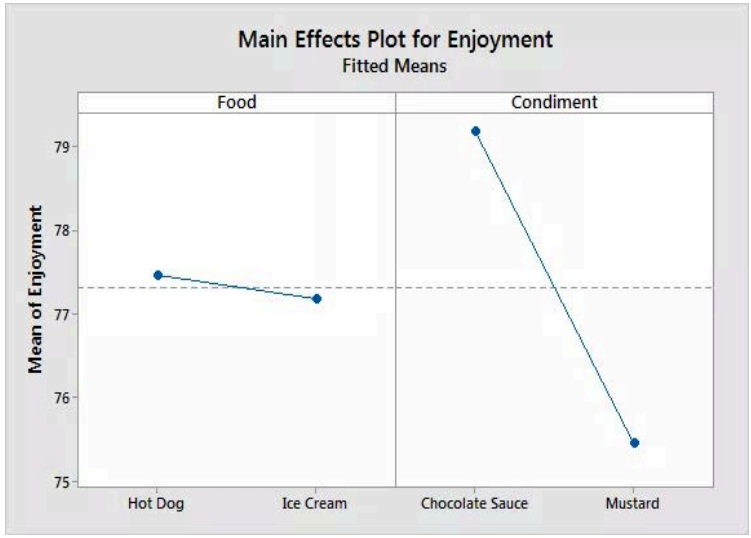
Condiments = People like chocolate sauce more than mustard



Suppose we want to maximize satisfaction by choosing the best food and the best condiment

But we forgot to include the interaction effect and assessed only the main effects.

We'll make our decision based on the main effects plots below.



Based on the plots, we'd choose hot dogs with chocolate sauce because they each produce higher enjoyment

https://metro.co.uk/2016/07/04/hot-dog-ice-cream-is-now-a-thing-that-exists-in-the-world-5983358/

Perhaps that is not the best despite what the main effects show!

When you have statistically significant interactions, you cannot interpret the main effect without considering the interaction effects

Always verify the interaction!

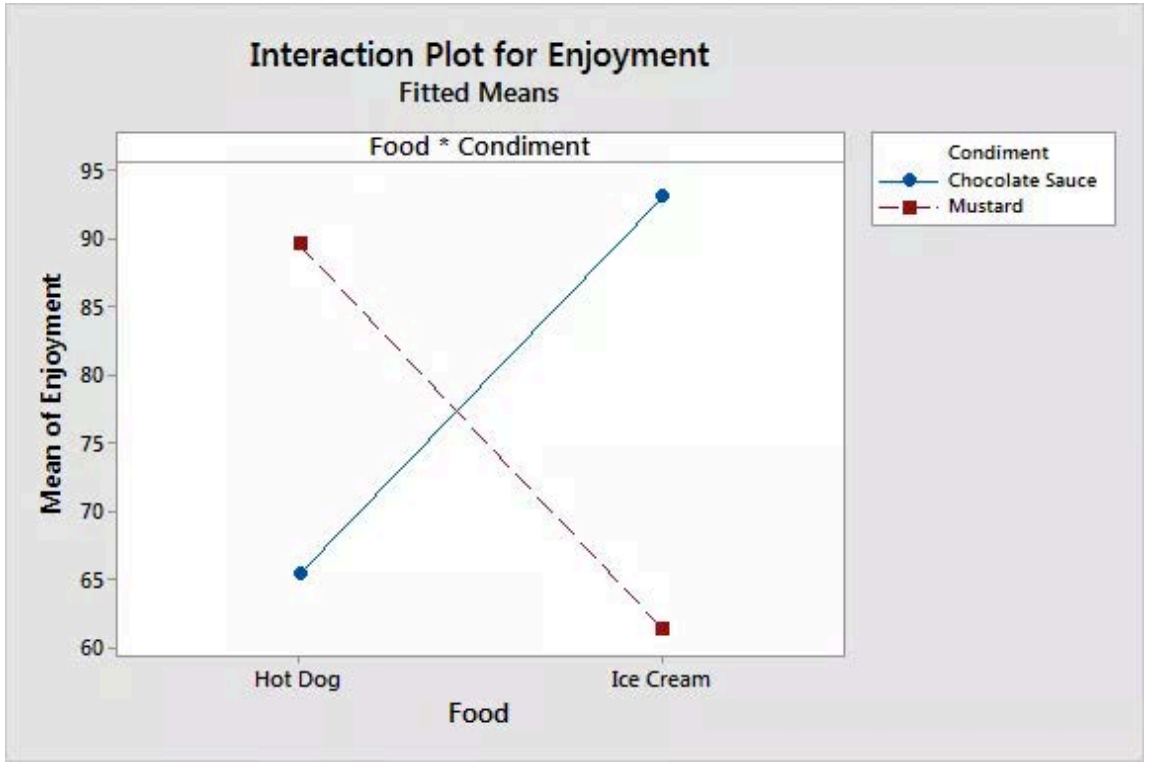
- http://statisticsbyjim.com/regression/interaction-effects/

Factor Information

Factor	Type	Levels	Values
Food	Fixed	2	Hot Dog, Ice Cream
Condiment	Fixed	2	Chocolate Sauce, Mustard

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Food	1	1.6	1.6	0.06	0.801
Condiment	1	277.5	277.5	11.07	0.001
Food*Condiment	1	15695.8	15695.8	626.15	0.000
Error	76	1905.1	25.1		
Total	79	17880.0			



How?

Two-Way ANOVA with interaction

```
aov(y ~ factor1*factor2)
```

a. If Interaction significant STOP and interpret results

b. If Interaction non significant eliminate interaction

Two-Way ANOVA without interaction

```
aov(y ~ factor1 + factor 2)
```

Two-way ANOVA with R

1. Download the file two_way.csv to your working directory for day 3

2. Open

```
#read data
two <- read.csv("two_way.csv")
head(two)

#Independent variables as factors
class(two$Temp)
class(two$Nutr)

two$Temp <- as.factor(two$Tem)
two$Nutr <- as.factor(two$Nutr)
```

```
# One-Way ANOVA Factor1 (temp)

f1 <- aov(Y ~ Temp, data = two)
summary(f1)

#p=0.07 p=fail to reject Ho

plot(TukeyHSD(f1))

# One-Way ANOVA Factor2 (Nutr)
f2 <- aov(Y ~ Nutr, data = two)
summary(f2)
#p=0.6- failt to reject Ho
plot(TukeyHSD(f2))
```

```
#Two-Way ANOVA interaction

f1f2int <- aov( Y ~ Temp * Nutr, data = two)

summary(f1f2int)

# ALL p-values <0.05!!
```

To determine the size of the interaction use the Eta square

The resulting eta squared value according to Cohen's (1988) terms would be considered:

.01 as a small effect, .06 as a medium effect and above .14 as a large effect.

```
library(DescTools)
EtaSq(f1f2int)
```

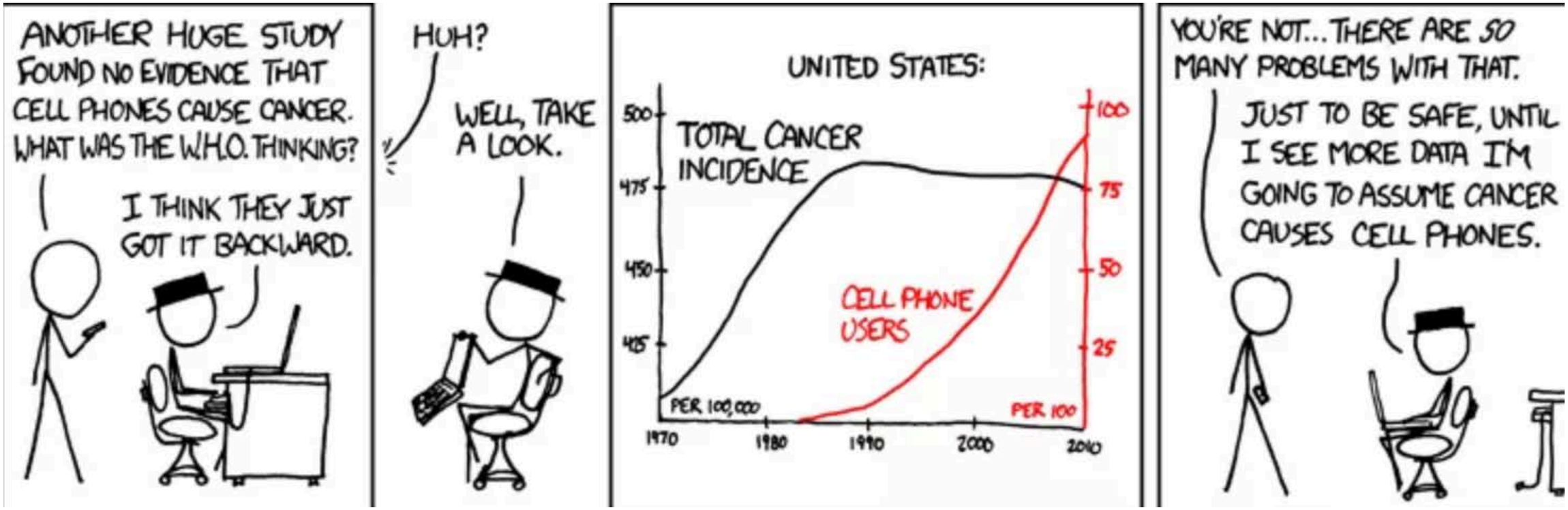
```
> EtaSq(f1f2int)
               eta.sq eta.sq.part
Temp       0.3064481   0.6667193
Nutr       0.2878799   0.6526887
Temp:Nutr  0.2524842   0.6223851
```

One-way and Two-way ANOVA summary

	One-way ANOVA	Two-way ANOVA
Definition	Compare between the **means of three or more groups** of data	Compare between the **means of three or more groups** of data, where **two independent variables** are considered
Number of independent variables	1	2
What is being compare	The means of three or more groups of an independent variable on a dependent variable	The effect of multiple groups of two independent variables on a dependent variable and on each other
Number of group of samples	The effect of multiple groups of two independent variables on a dependent variable and on each other	Each variable should have multiple samples

https://www.technologynetworks.com/informatics/articles/one-way-vs-two-way-anova-definition-differences-assumptions-and-hypotheses-306553

Correlation



https://i2.wp.com/www.skepticalraptor.com/blog/wp-content/uploads/2014/05/cell-phones-cancer.png

Statistical inference for the relationship between two variables

Correlation analysis is used to describe the strength and direction of the linear relationship between two variables

Statistical inference for the relationship between two variables

There are number of statistics, e.g.

- Pearson product-moment coefficient: for continuous variables
- Spearman rank order correlation: use with ordinal level or ranked data

Pearson Correlation

- Pearson correlation coefficients (r) can take on only values from –1 to +1
- The sign indicates whether there is a positive correlation (as a variable increases, so too does the other) or a negative correlation (as one variable increases, the other decreases)

Pearson Correlation

- The size of the absolute value (ignoring the sign) provides an indication of the strength of the relationship
- A perfect correlation of 1 or –1 indicates that the value of one variable can be determined exactly by knowing the value on the other variable

Correlation Analysis with R

1. Download the file correlations.csv to your working directory for day 3

2. Open

```
cor.test(Corr1$SST, Corr1$production_mgCm2d,
alternative = c("two.sided", "less", "greater"))


cor.test(Corr1$pH, Corr1$production_mgCm2d,
alternative = c("two.sided", "less", "greater"))
```

Correlation between PP and SST

```
        Pearson's product-moment correlation

data:  Corr1$SST and Corr1$production_mgCm2d
t = -21.736, df = 34, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9825932 -0.9335431
sample estimates:
       cor
-0.9658503
```

Correlation between PP and SST

```
        Pearson's product-moment correlation

data:  Corr1$SST and Corr1$production_mgCm2d
t = -21.736, df = 34, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9825932 -0.9335431
sample estimates:
       cor
-0.9658503
```

Strong evidence against Ho (rho=0),therefore we reject Ho.

Correlation between PP and pH

```
        Pearson's product-moment correlation

data:  Corr1$pH and Corr1$production_mgCm2d
t = 1.3237, df = 34, p-value = 0.1944
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1155577  0.5126336
sample estimates:
      cor
0.2213828
```

Correlation between PP and pH

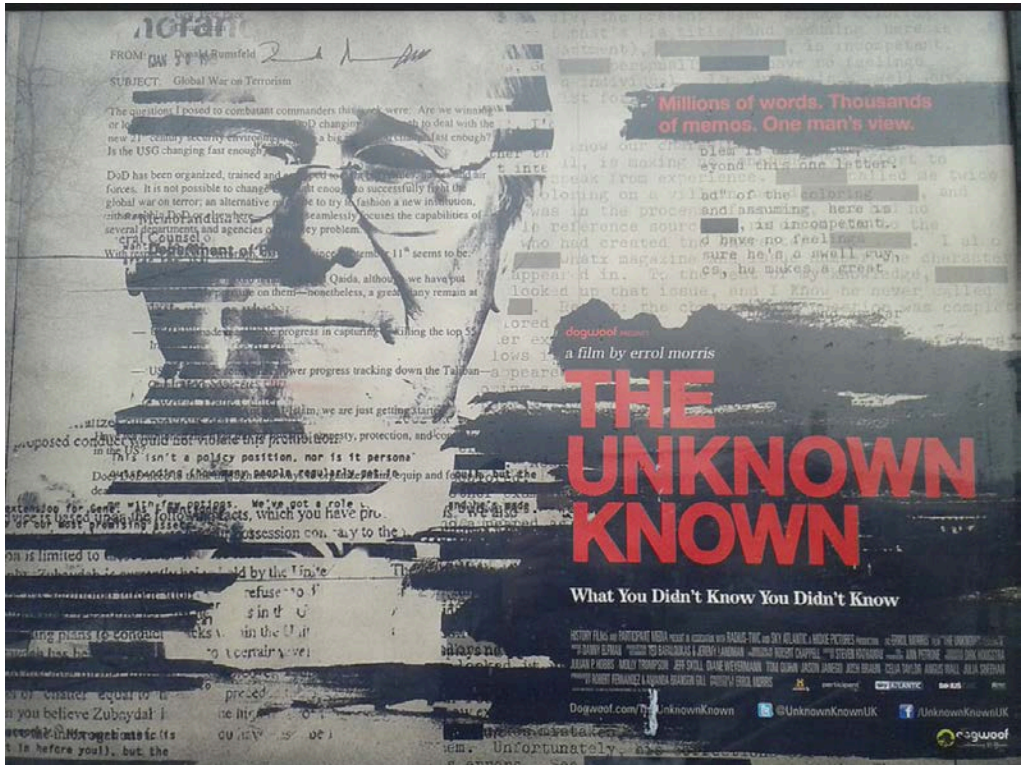
```
        Pearson's product-moment correlation

data:  Corr1$pH and Corr1$production_mgCm2d
t = 1.3237, df = 34, p-value = 0.1944
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1155577  0.5126336
sample estimates:
      cor
0.2213828
```

Week evidence against Ho (rho=0),therefore we rail to reject Ho.

Correlation Matrix with R

1. Let’s go back to 

Regression Analysis



https://www.flickr.com/photos/dullhunk/13569655024

Regression Analysis

We use linear regression models for either **testing a hypothesis regarding the relationship** between one or more explanatory variables and a response variable, or **predicting unknown values of the response variable** using one or more predictors

Regression Analysis

The regression line captures the linear relationship between the response variable and the explanatory variable

Rasmus Bäät

https://www.r-bloggers.com/the-most-comprehensive-review-of-comic-books-teaching-statistics/

Model Assumptions

Linearity:

The relationship between the explanatory variable x and the response variable y is linear

If no linearity then you may transform, common transformations are:

- Logarithmic (usually for the response variable)
- Square root and square (usually for predictors)

Model Assumptions

Independence:

Observations must be independent, i.e., simple random sampling to select individuals that are not related to each other and not multiple observations from the same individual

Model Assumptions

Constant variance and normality: minor deviation from normality will not have a significant impact on the results as long as the sample size is relatively large

Model Assumptions

Constant variance and normality:

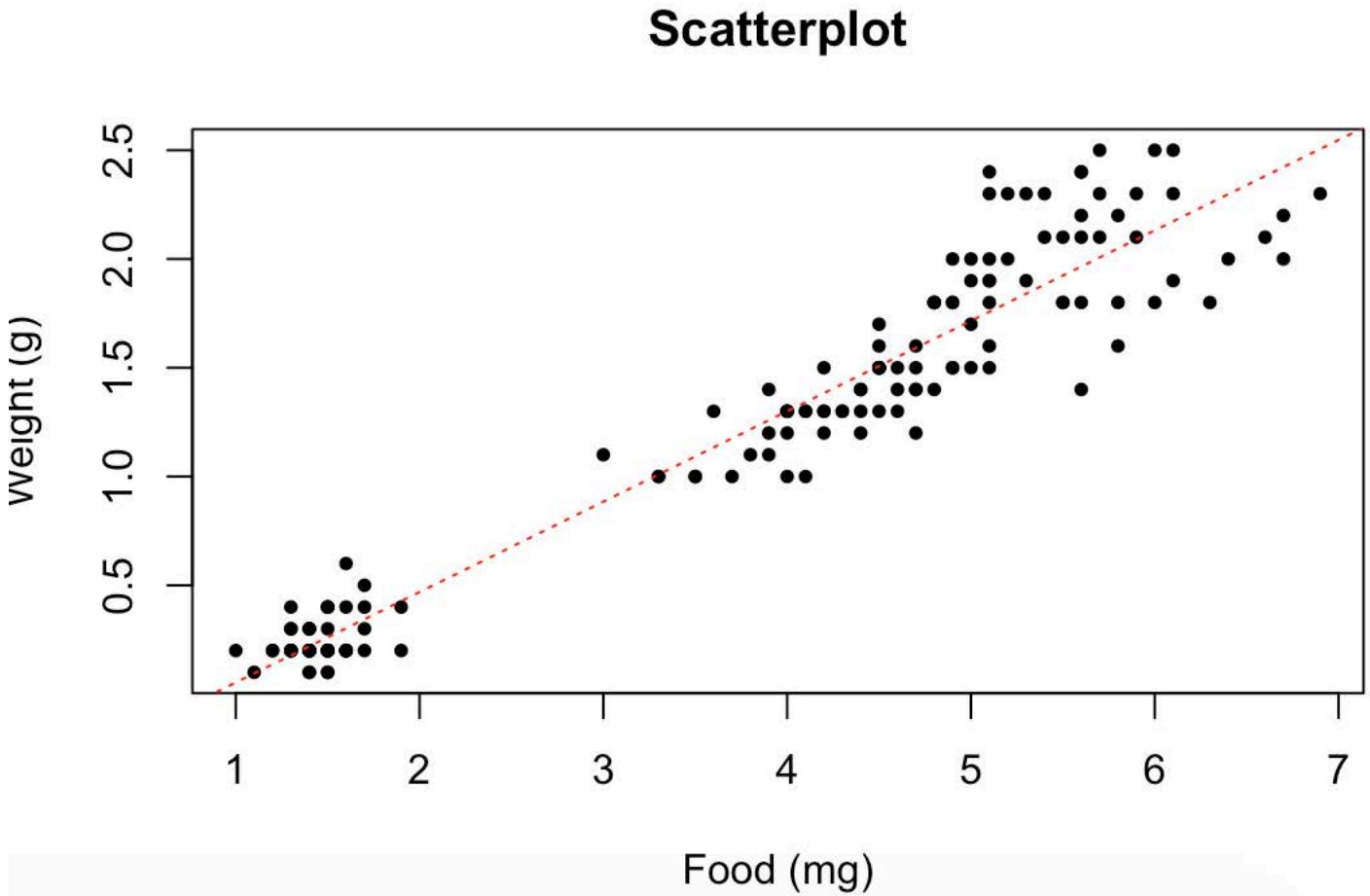
We expect that the **variation** of the actual values **of the response variable** around the regression line **remains** the same **regardless of the value of the explanatory variable.**

This is called the constant variance assumption, which is also known as homoscedasticity assumption

Regression Analysis with R

1. Download the file Regresions.csv to your working directory for day 3
2. Open

```
Lmodel2 <- lm(Reg1$Weigh_g ~ Reg1$mg_food)
print(Lmodel2)
summary(Lmodel2)

plot(Reg1$mg_food, Reg1$Weigh_g,
main="Scatterplot",
     xlab="Food (mg)", ylab="Weight (g)",
pch=19,cex=0.6) # line (x,y)
abline(Lmodel1, col="blue", lty=2)

pred2 <- c(1, 5, 10)
coef(Lmodel2)[1] + coef(Lmodel2)[2] * pred2
```


Scatterplot
Weight (g)
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7
Food (mg)

Regression Analysis

```
> print(Lmodel2)

Call:
lm(formula = Reg1$Weigh_g ~ Reg1$mg_food)

Coefficients:
 (Intercept)  Reg1$mg_food
     -0.3626        0.4157
```

Regression Analysis

```
Coefficients:
 (Intercept)  Reg1$mg_food
     -0.3626        0.4157
```

We estimate an expected **0.41 g increase in weight** for **every mg food** added

The intercept -0.3636 is the expected increase in weight if we add 0 g of food

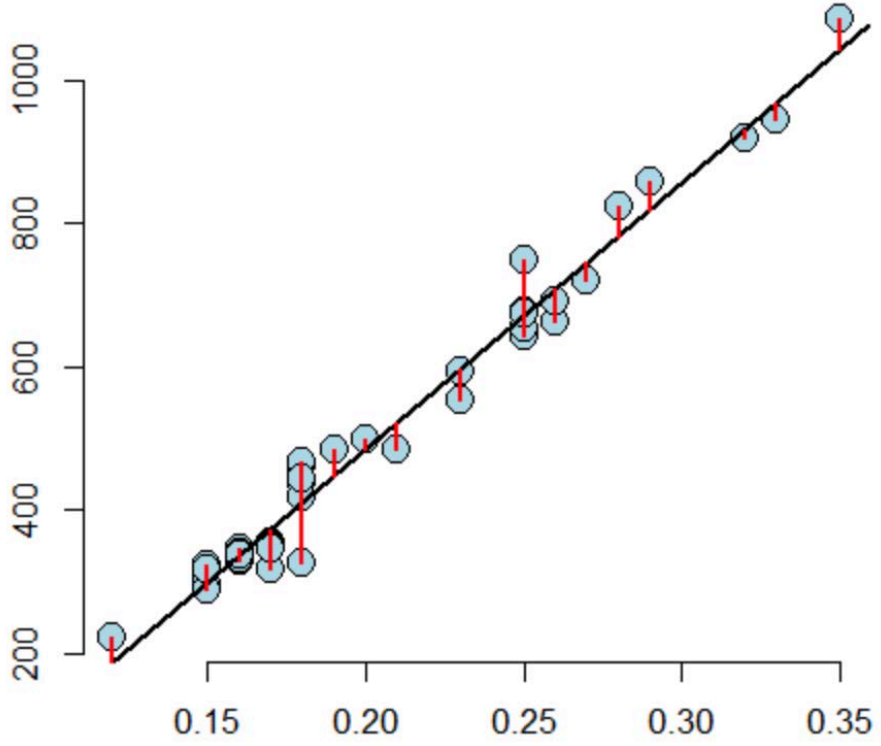
```
> summary(Lmodel2)

Call:
lm(formula = Reg1$Weigh_g ~ Reg1$mg_food)

Residuals:
      Min        1Q    Median        3Q       Max
-0.56510 -0.12354 -0.01934  0.13490  0.64273
```

Residuals represent variation left unexplained by our model

So they are observable errors from the estimated coefficients

Hence, the residuals are estimates of the errors

The residuals are exactly the vertical distance between the observed data point and the associated point on the regression line



```
> summary(Lmodel2)

Call:
lm(formula = Reg1$Weigh_g ~ Reg1$mg_food)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56510 -0.12354 -0.01934  0.13490  0.64273

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.362577   0.040258  -9.006 1.02e-15 ***
Reg1$mg_food  0.415657   0.009673  42.970  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2072 on 147 degrees of freedom
Multiple R-squared:  0.9263,    Adjusted R-squared:  0.9258
F-statistic:  1846 on 1 and 147 DF,  p-value: < 2.2e-16
```

```
Residual standard error: 0.2072 on 147 degrees of freedom
Multiple R-squared:  0.9263,    Adjusted R-squared:  0.9258
F-statistic:  1846 on 1 and 147 DF,  p-value: < 2.2e-16
```

The percentage of the total variability that is explained by the linear relationship with the predictor (R2) was 0.92

All good right?

...Not really!

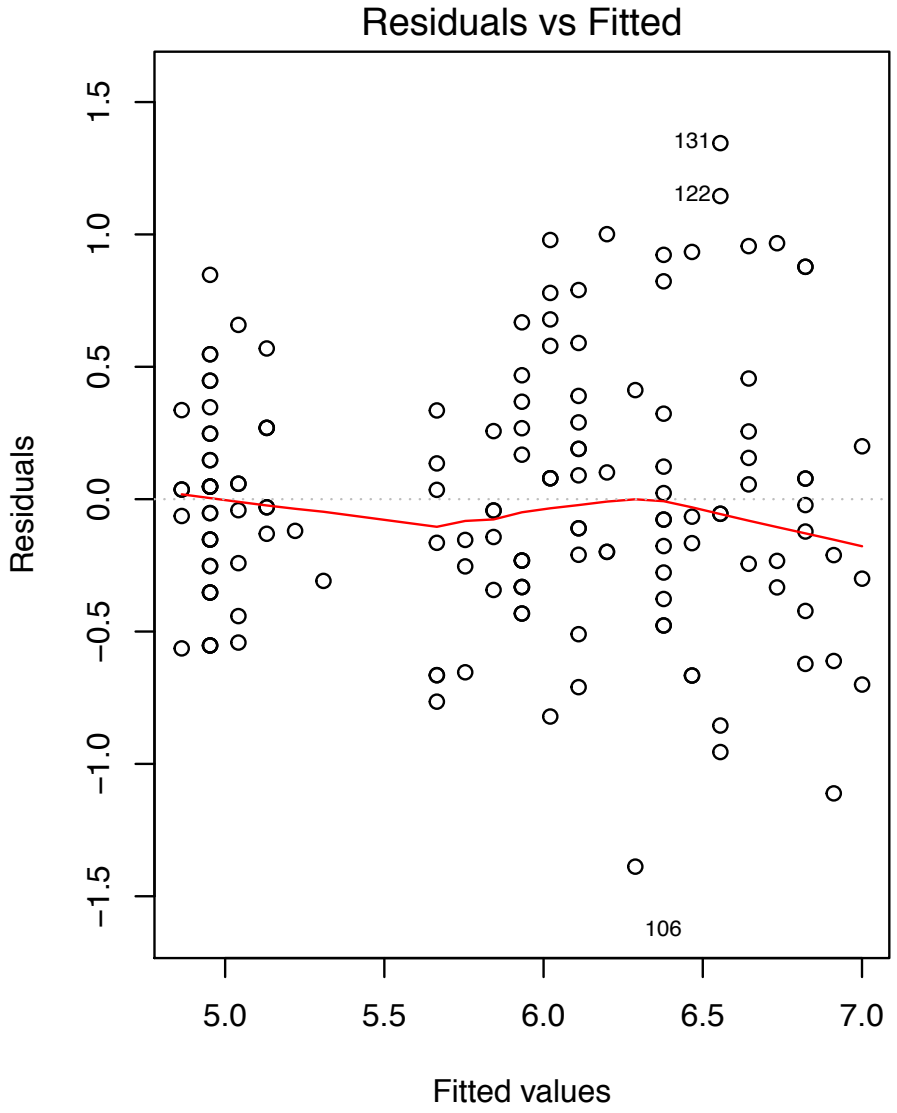
How about the assumptions?

Linear regression assumptions

- Linearity of the data
- Independent and random observations
- Normality of residuals
- Homogeneity of residuals variance (homoscedasticity)
- Independence of residual errors

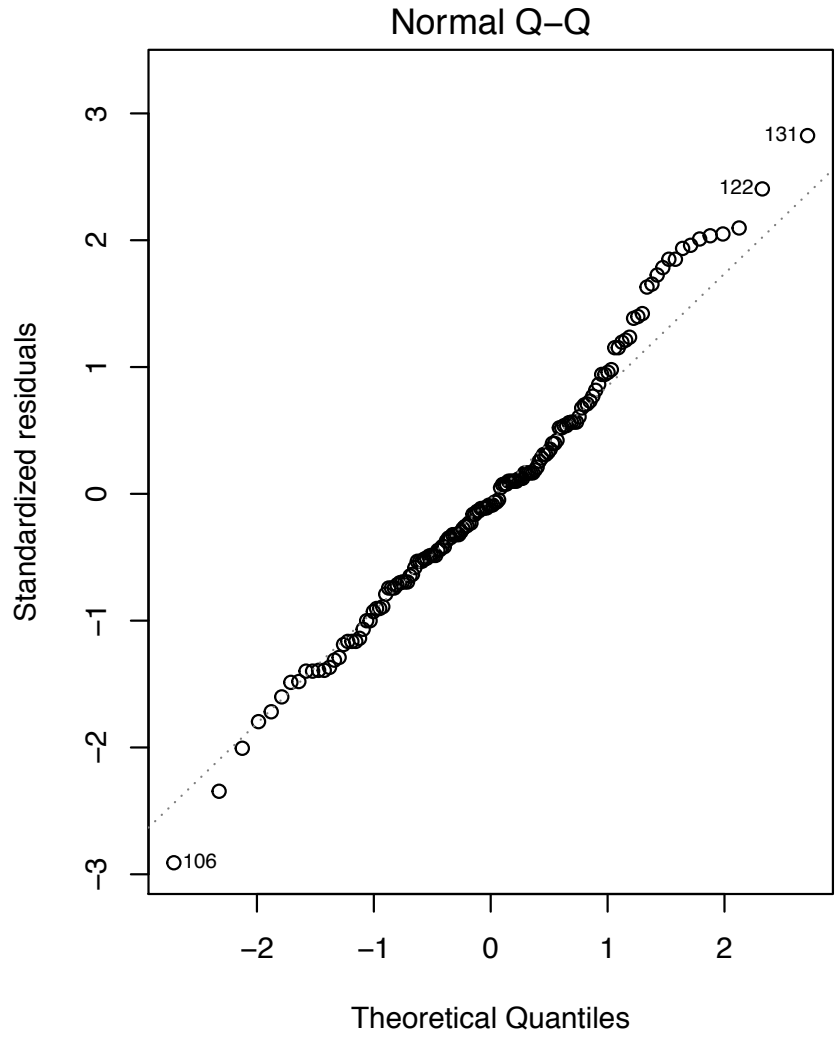
Linearity of the data

Ideally the plot won't show any fitted pattern, and the red line should be approx. to 0



Normality of the Residuals

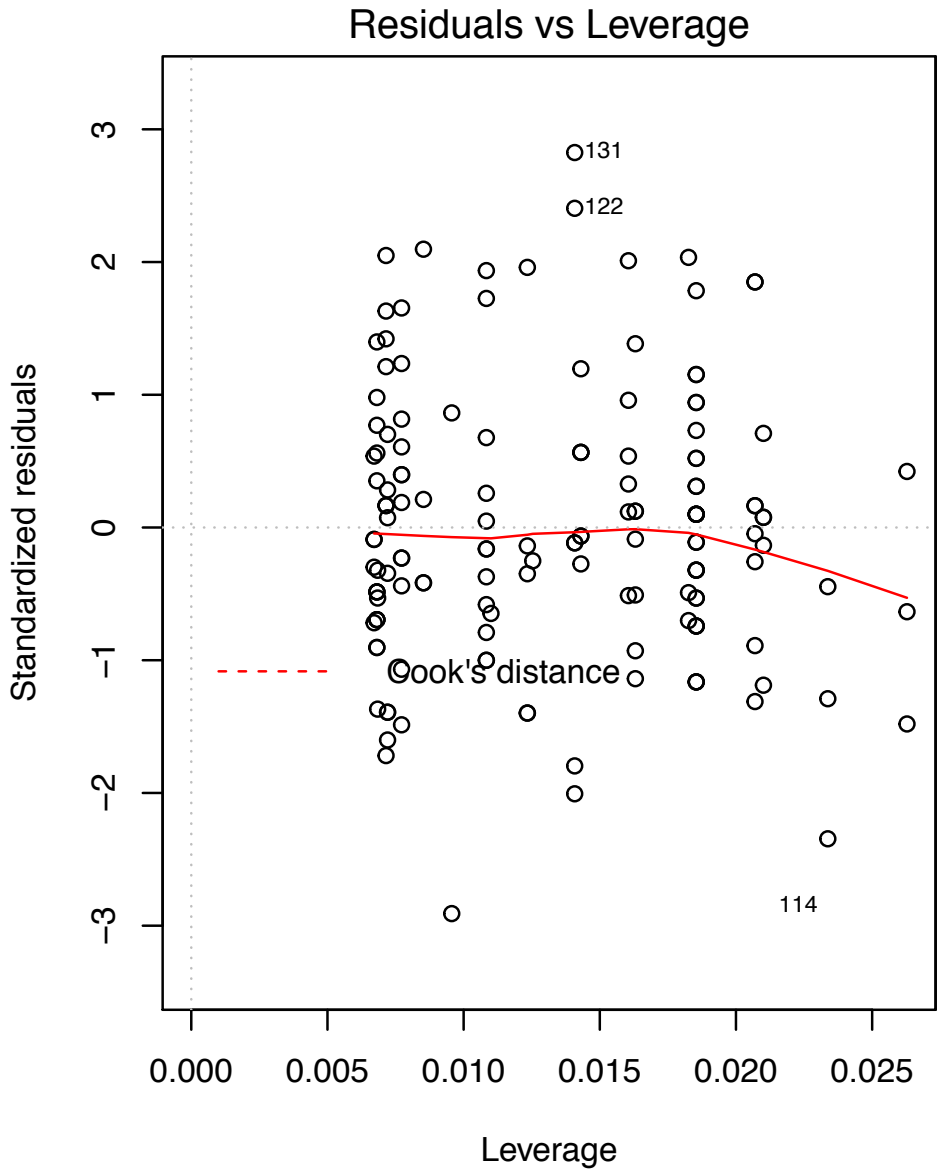
Should approximately follow a straight line



Influential points

Outliers: measured by examining the standardized residuals (SR)

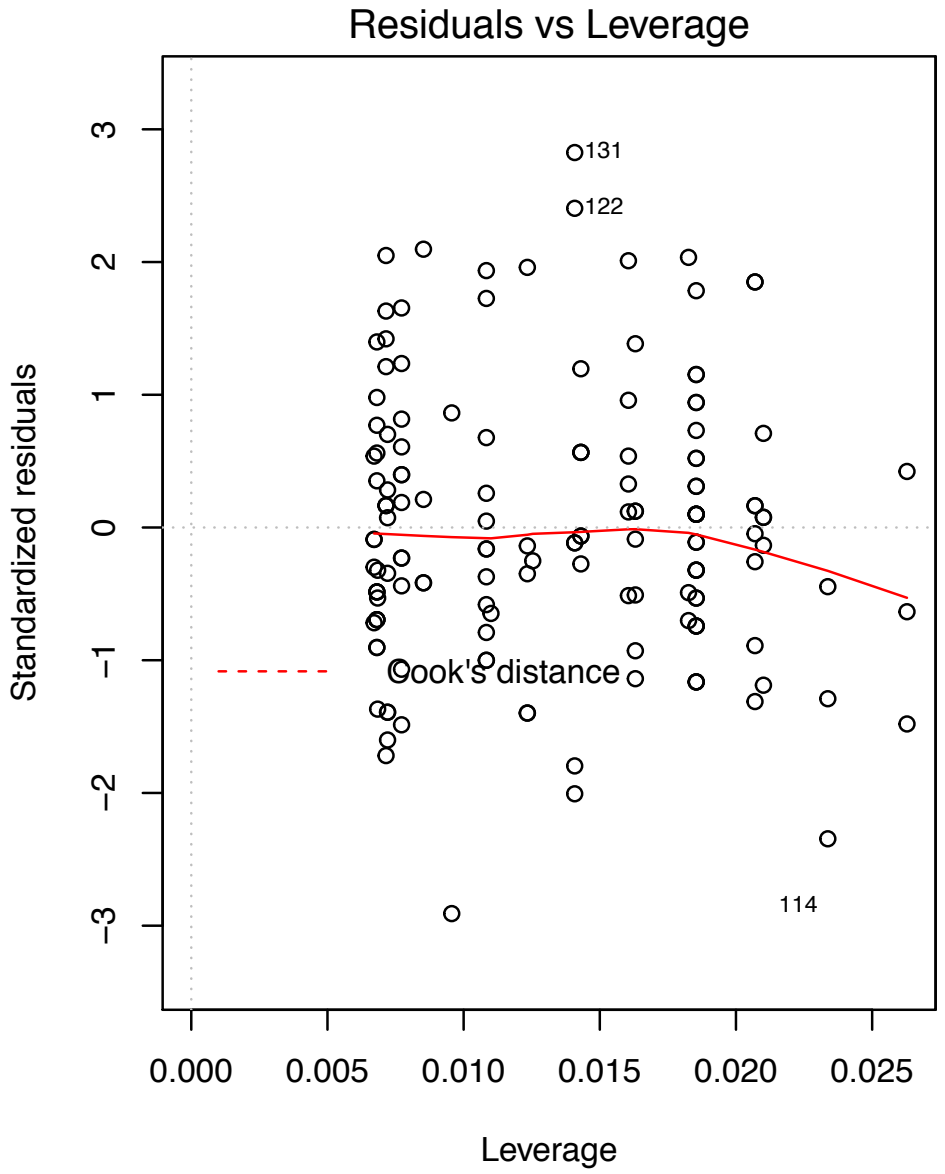
SR=Residuals/SE
Observations > 3 are possible outliers



Influential points

The leverage is a measure that describes the amount by which the predicted value would change if the observation was shifted on one unit in the y direction

Takes values between 0 and 1



Influential points

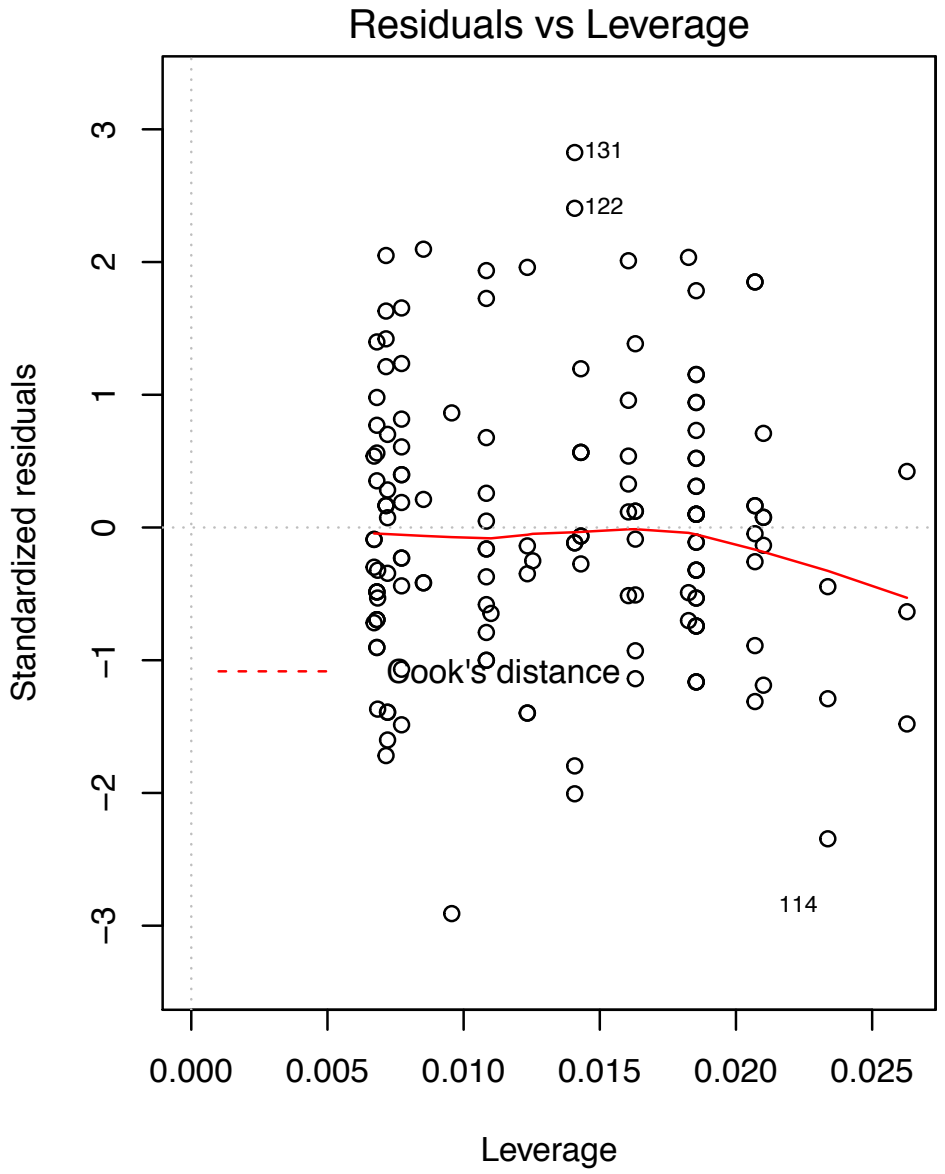
High leverage points:

A value above 2(p+1)/n indicates high leverage.

p = the number of predictors
n = the number of observations

2(1+1)/149 =0.027

No high leverage all values below 0.027

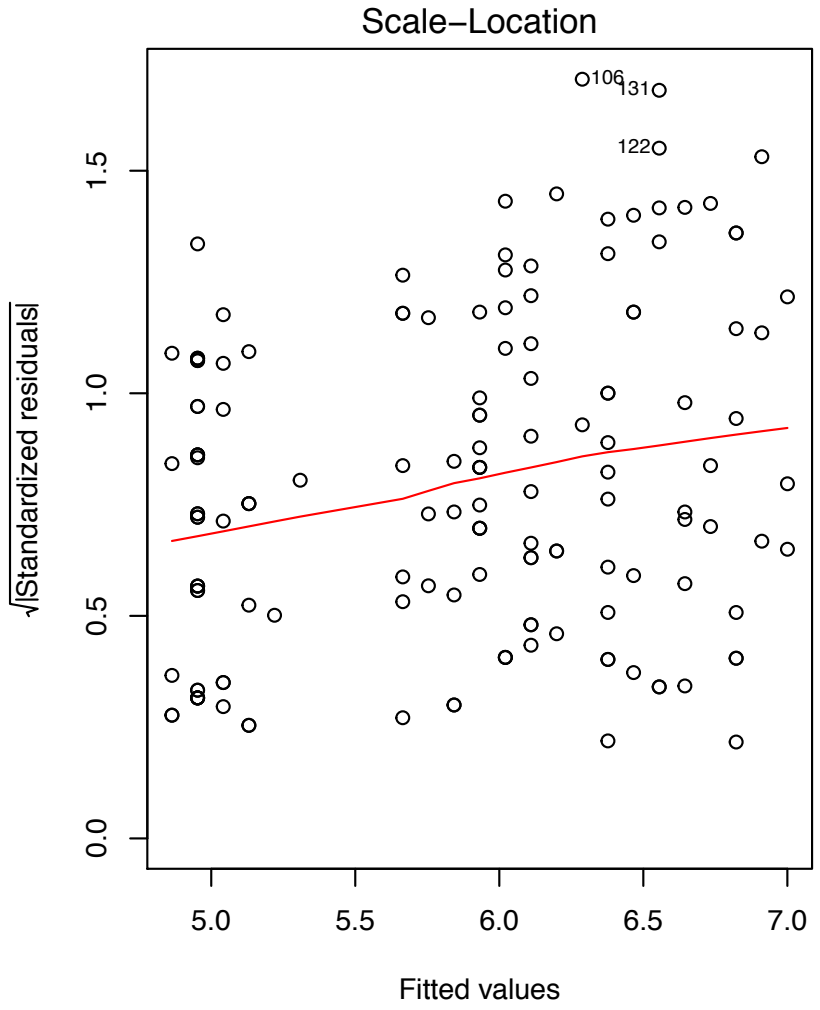


Homogeneity of variance

Determine if the residuals are equally along the ranges of predictors.

If you see a horizontal line with equally spread points (nor our case) we are good.

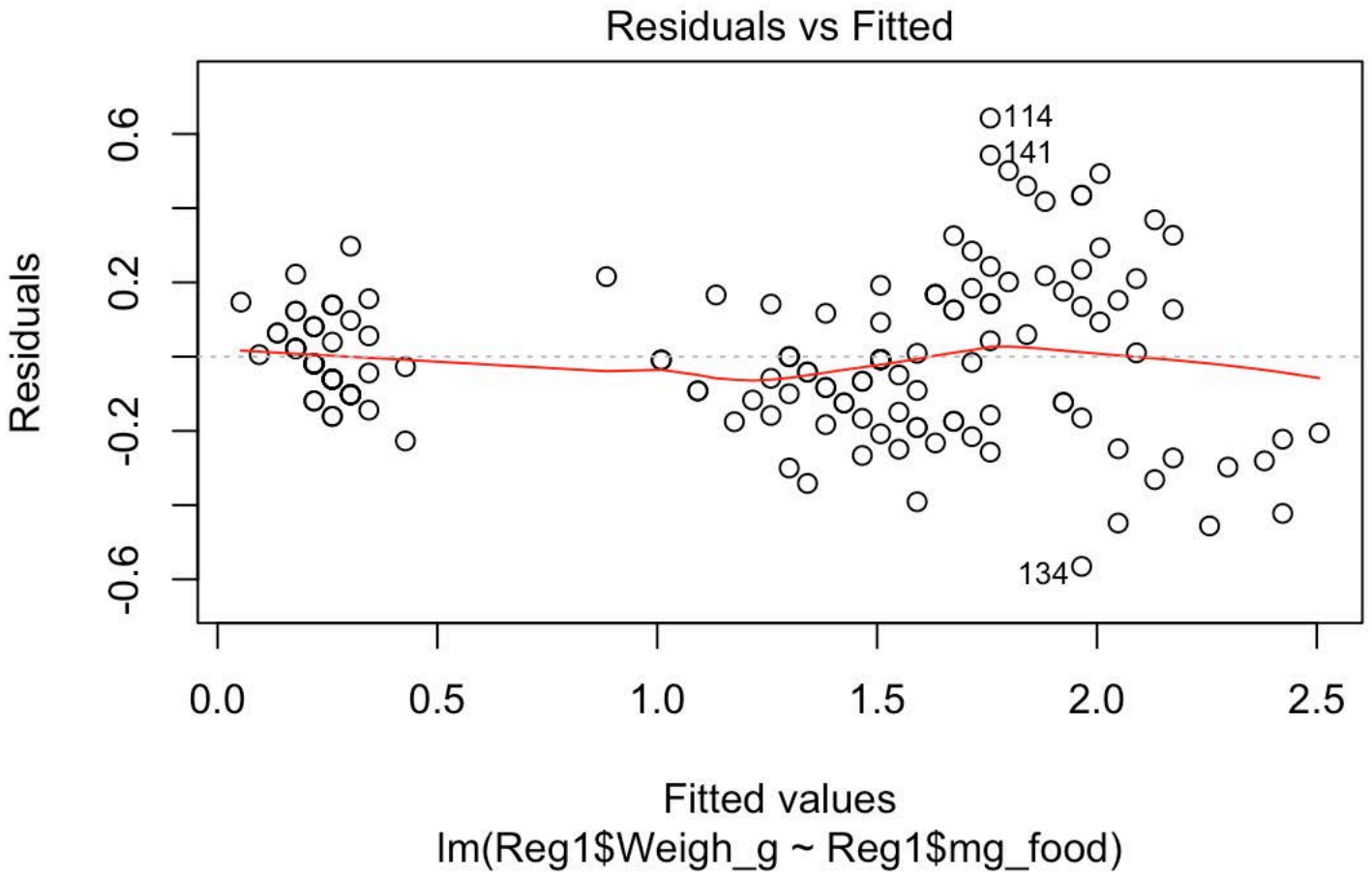
If not, a possible sol is to transform (log or sqr) of our outcome variable (y)



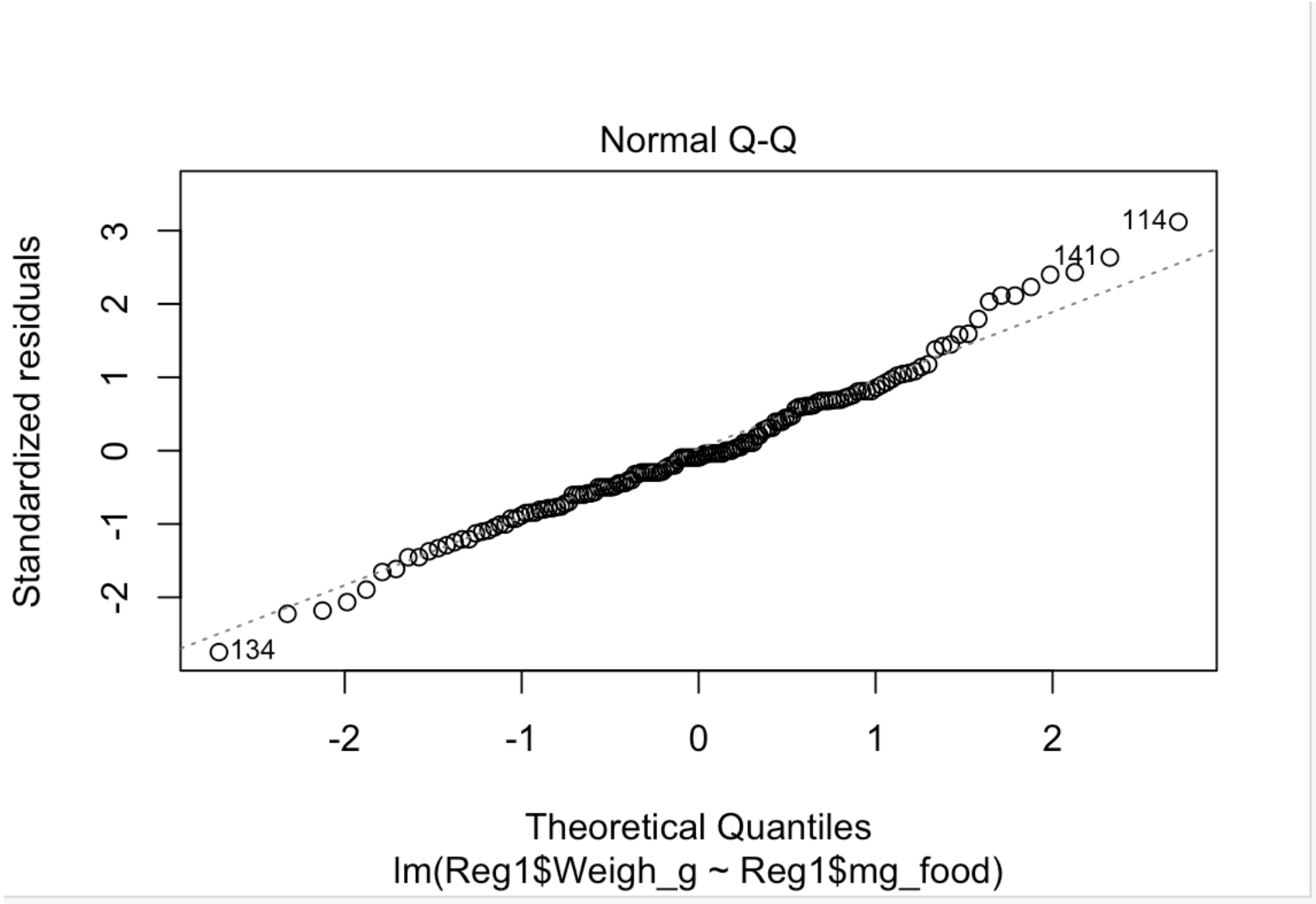
Let’s go back to our model and test assumptions

```
fit2 <- lm(Reg1$Weigh_g ~ Reg1$mg_food, data = Reg1)

summary(fit2)

plot(fit2)

# Calculate Leverage=
2*(p+1)/n= 2*(1+1)/149
[1]0.02684564
```

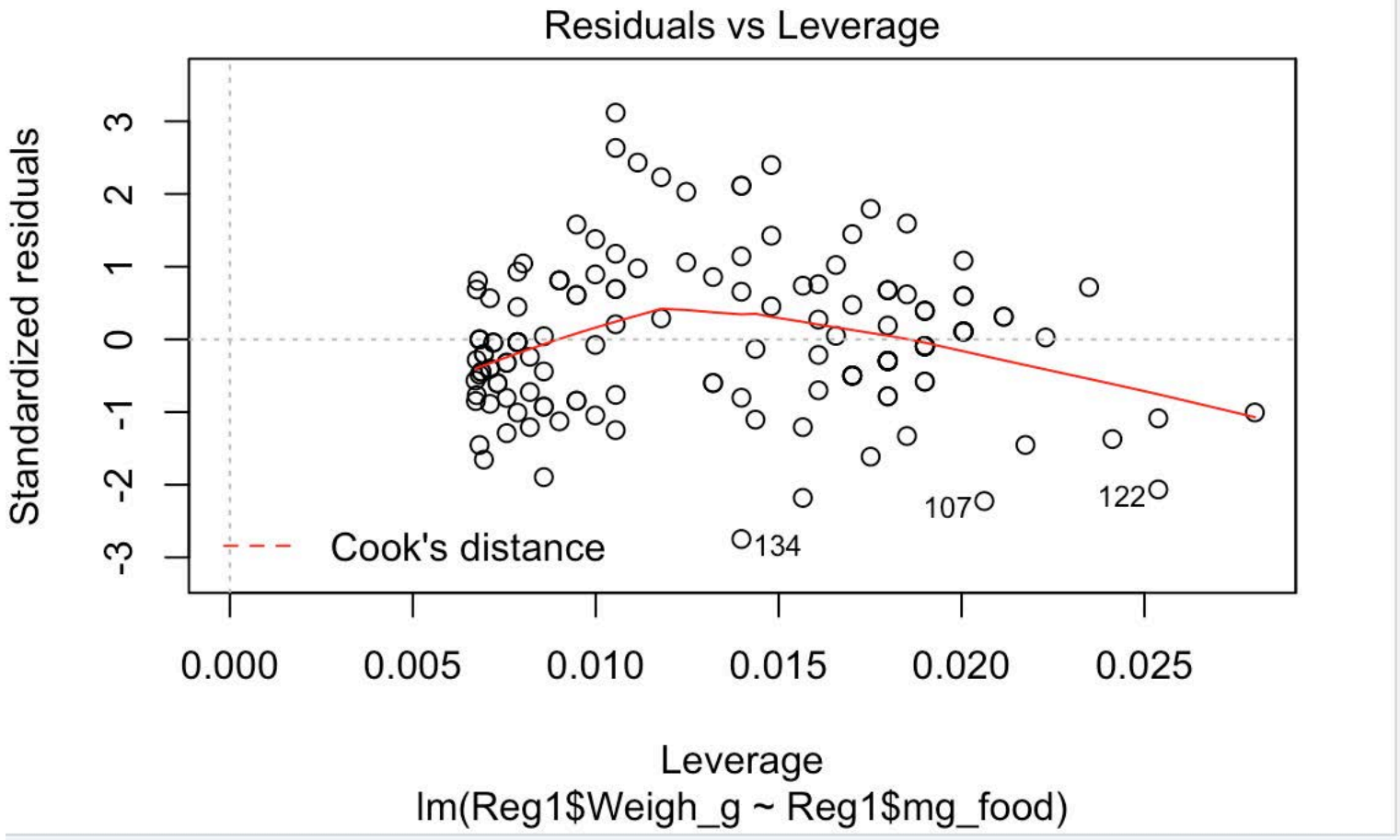
Linearity of the data



Normality of the Residuals

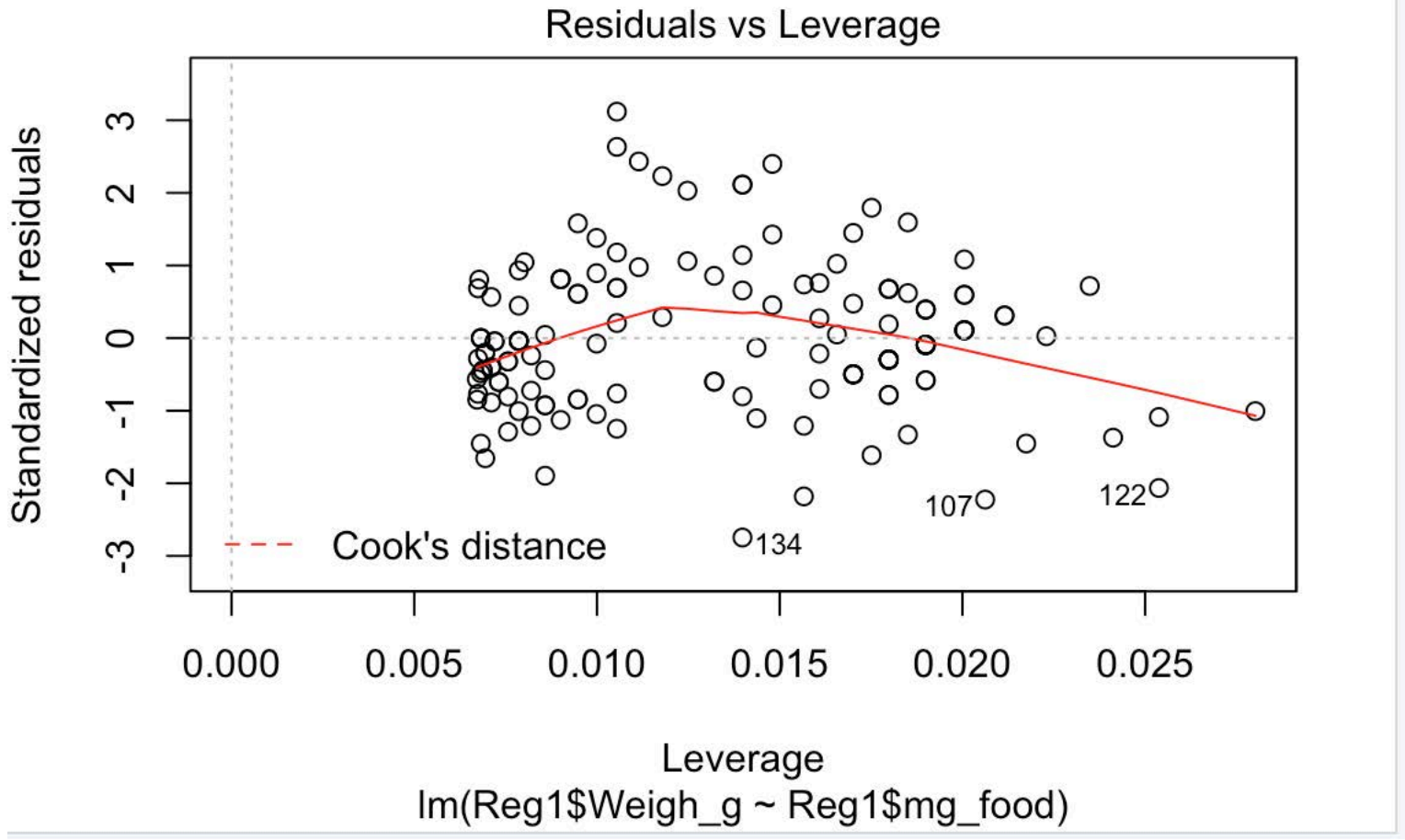


Influential points

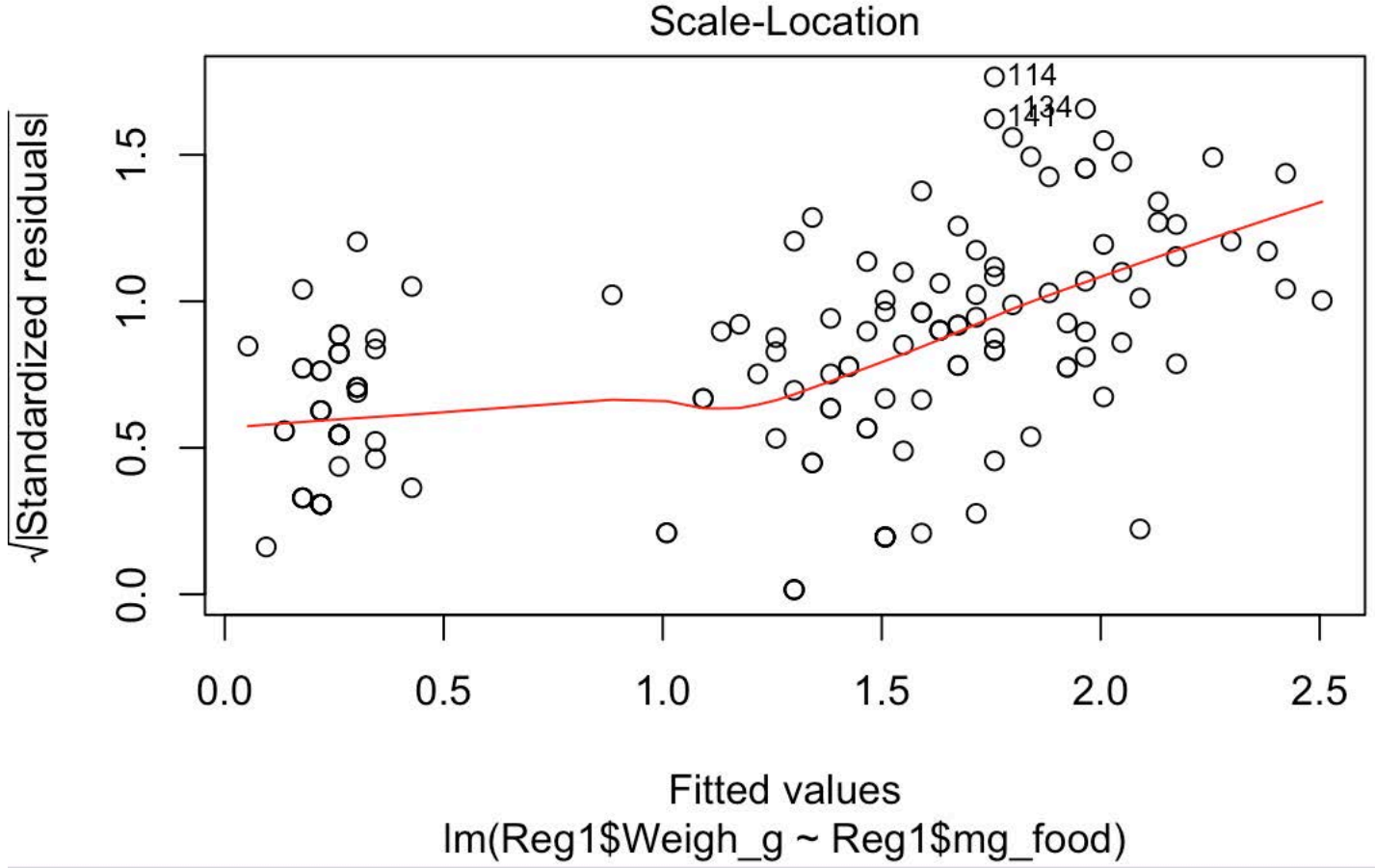


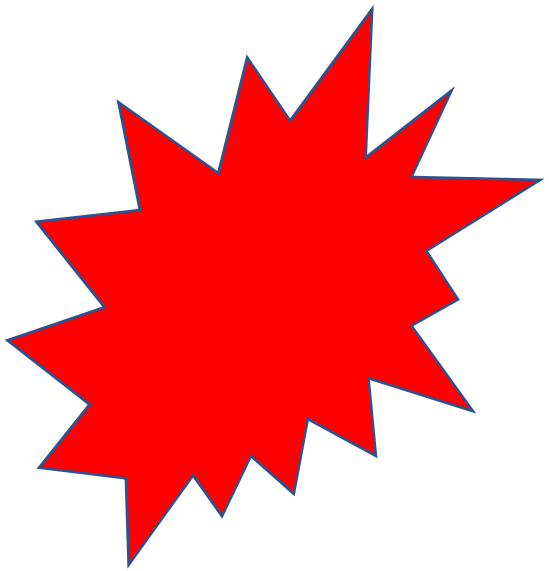
Influential points

Calculated leverage= 0.027



Equal Variances (Homoscedasticity)





```
## Residual analysis

sresid2 <- studres(fit)
hist(sresid2, freq=FALSE,
     main="Distribution of Studentized Residuals")

##Studentized Res=Residuals/SD

# normal dist residuals
shapiro.test(sresid2) #p=0.65, fail to reject Ho

#Homocedasticity
ncvTest(fit2)# p<0.05!
```