Introduction to statistics

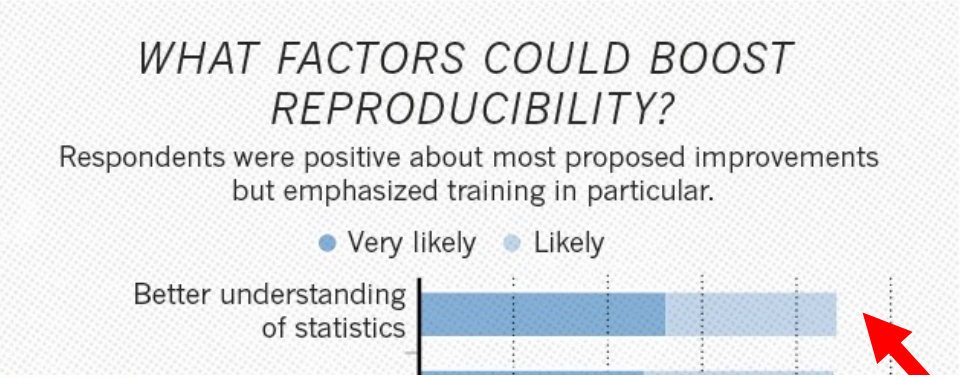
"There are three kind of lies: lies, damned lies and statistics." -Disraeli

Marco La Fortezza, ETH Zürich
marco.lafortezza@env.ethz.ch

When a life scientist faces statistics...

- **Handle** numerical data and visualize them
- **Test hypothesis** using statistics (not only the bloody p value)
- I **need** stars on my beautiful graphs!
- Look and **interpret** other people data
- **Design** your experiment to get sustainable and reproducible results!

The strange relationship between life science and statistics



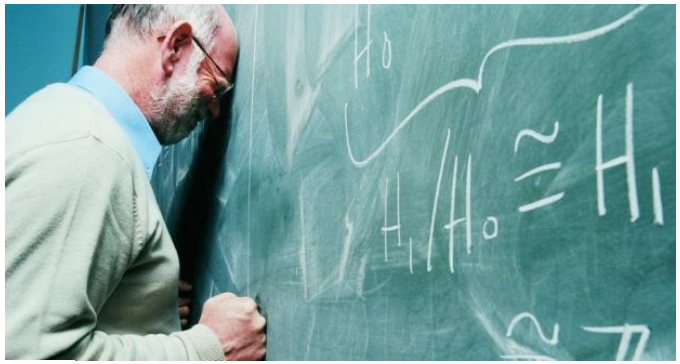


Table 1.

Design of statistics survey and student responses

Item	Category	Survey question	Correct responses presurvey (%)[a]	Correct responses postsurvey (%)[a]
1	Identification	Histogram	67 (36)	127 (69)
2	Identification	Scatterplot	183 (98)	184 (99)
3	Identification	Stem-and-leaf plot	149 (80)	184 (99)
4	Graphical tools	Box plot	113 (61)	142 (76)
5	Graphical tools	Histogram	99 (53)	116 (62)
6	Graphical tools	Scatterplot	140 (75)	157 (84)
7	Experimental analysis	Quartiles	162 (88)	175 (94)
8	Experimental analysis	*p* value	66 (35)	128 (69)
9	Experimental analysis	Causation vs. correlation	96 (52)	126 (68)
10	Data evaluation	Causation/correlation	87 (47)	90 (48)
11	Data evaluation[b]	*p* value	65 (35)	81 (43)

[a] Total students participating in survey n = 186.

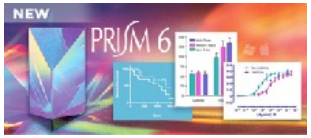
[b] Includes both answer choices C and E.

CBE Life Sci Educ

A.M. Metz, CBE Life Science Education 2008

M. Baker, Nature 2016

Where to run statistics



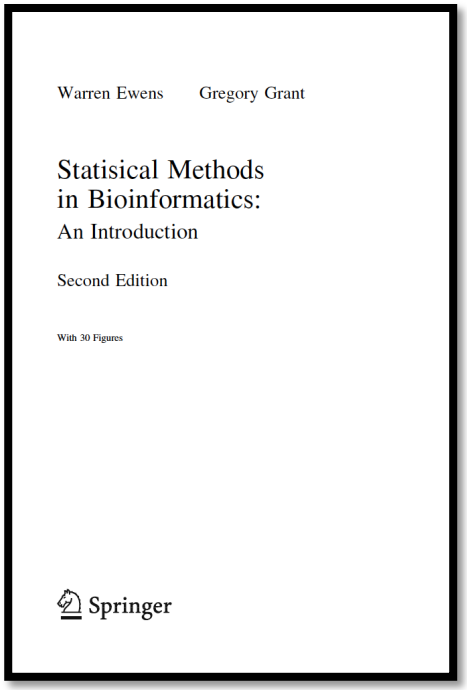
	EXCEL	Prism	R
Price	MEDIUM	**HIGH**	**FREE**
Ease of use	MEDIUM	**EASY**	**DIFFICULT**
Versatility	**LOW**	HIGH	**YOU CAN DO ANYTHING**
Misusage	MEDIUM	**EASY**	MEDIUM
Data Visualization	**POOR**	GOOD	**BEST**

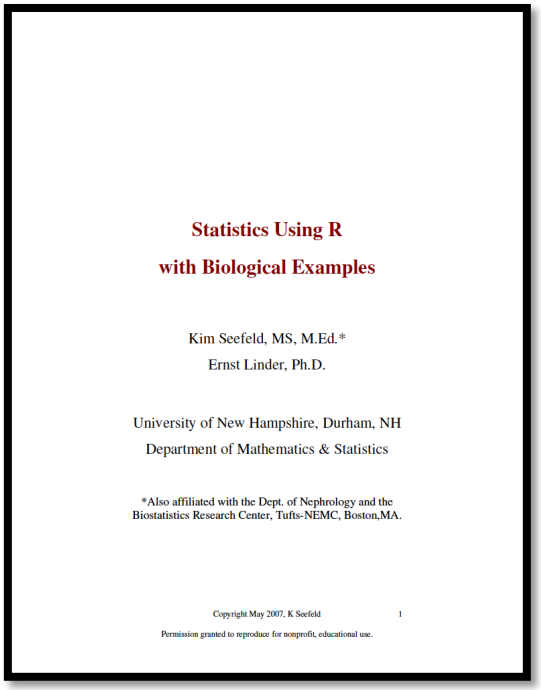
Some resources

The Analysis of Biological Data – Whitlock, Schluter
http://whitlockschluter.zoology.ubc.ca/

Handbook of Biological Statistics – J.H. McDonald
http://www.biostathandbook.com/index.html

Statistics for biologist - Nature
http://www.nature.com/collections/qghhqm





Type of different biological variables

Armadillium vulgare

You want to measure a **variable X** of *Armadillidium vulgare* in 56 males and 67 females. You want to check whether the **variable X is the same in males and females.**

- **Measurements variables**

  if you are measuring the length of the animals (continuous scale)

- **Nominal variables**

  if are comparing the genotype (aa, Aa, AA) of males and females

- **Ranked variables**

  if you measure the speed of the two groups of isopod to unroll

How to collect data

Variables

Observations, Samples

	A	B	C	D	E
1	year	sex	oceanAgeYears	lengthMm	massKg
2	1996	F	2	447	2.029
3	1996	F	2	437	2.033
4	1996	F	2	447	2.036
5	1996	F	2	447	2.068
6	1996	F	2	447	2.118
7	1996	F	2	457	2.13
8	1996	F	2	457	2.132
9	1996	F	2	457	2.137
10	1996	F	2	457	2.156
11	1996	F	2	447	2.177
12	1996	F	2	457	2.203
13	1996	F	2	468	2.259
14	1996	F	2	457	2.316
15	1996	F	2	468	2.362
16	1996	F	2	468	2.615
17	1996	F	3	457	1.941
18	1996	F	3	479	2.288
19	1996	F	3	468	2.334
20	1996	F	3	479	2.393
21	1996	F	3	468	2.425
22	1996	F	3	479	2.452
23	1996	F	3	479	2.495
24	1996	F	3	468	2.56
25	1996	F	3	501	2.572
26	1996	F	3	479	2.578
27	1996	F	3	490	2.59
28	1996	F	3	501	2.658
29	1996	F	3	501	2.677
30	1996	F	3	501	2.69
31	1996	F	3	490	2.691
32	1996	F	3	501	2.703
33	1996	F	3	513	2.753
34	1996	F	3	501	2.764
35	1996	F	3	501	2.768
36	1996	F	3	501	2.775
37	1996	F	3	501	2.808
38	1996	F	3	513	2.815
39	1996	F	3	501	2.834

Proper **data organization** can have a strong impact in **data exploration** and **data analysis**

Types of data

New Open Save Print Import Copy Paste Format Undo Redo AutoSum Sort A-Z

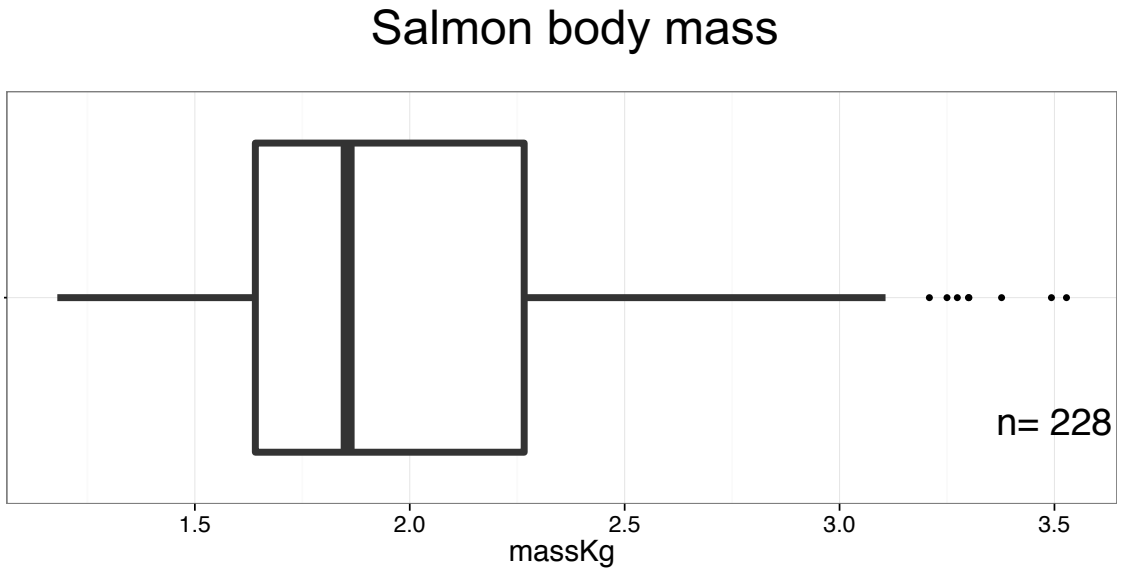
	A	B	C	D	E
1	year	sex	oceanAgeYears	lengthMm	massKg
2	1996	F	2	447	2.029
3	1996	F	2	437	2.033
4	1996	F	2	447	2.036
5	1996	F	2	447	2.068
6	1996	F	2	447	2.118
7	1996	F	2	457	2.13
8	1996	F	2	457	2.132
9	1996	F	2	457	2.137
10	1996	F	2	457	2.156
11	1996	F	2	447	2.177
12	1996	F	2	457	2.203
13	1996	F	2	468	2.259
14	1996	F	2	457	2.316
15	1996	F	2	468	2.362
16	1996	F	2	468	2.615
17	1996	F	3	457	1.941
18	1996	F	3	479	2.288
19	1996	F	3	468	2.334
20	1996	F	3	479	2.393
21	1996	F	3	468	2.425
22	1996	F	3	479	2.452
23	1996	F	3	479	2.495
24	1996	F	3	468	2.56
25	1996	F	3	501	2.572
26	1996	F	3	479	2.578
27	1996	F	3	490	2.59
28	1996	F	3	501	2.658
29	1996	F	3	501	2.677
30	1996	F	3	501	2.69
31	1996	F	3	490	2.691
32	1996	F	3	501	2.703
33	1996	F	3	513	2.753
34	1996	F	3	501	2.764
35	1996	F	3	501	2.768
36	1996	F	3	501	2.775
37	1996	F	3	501	2.808
38	1996	F	3	513	2.815
39	1996	F	3	501	2.834
40	1996	F	3	490	2.872
41	1996	F	3	513	2.872
42	1996	F	3	513	2.876

- discrete
- nominal
- continuous

Data Representation

Plotting continuous data

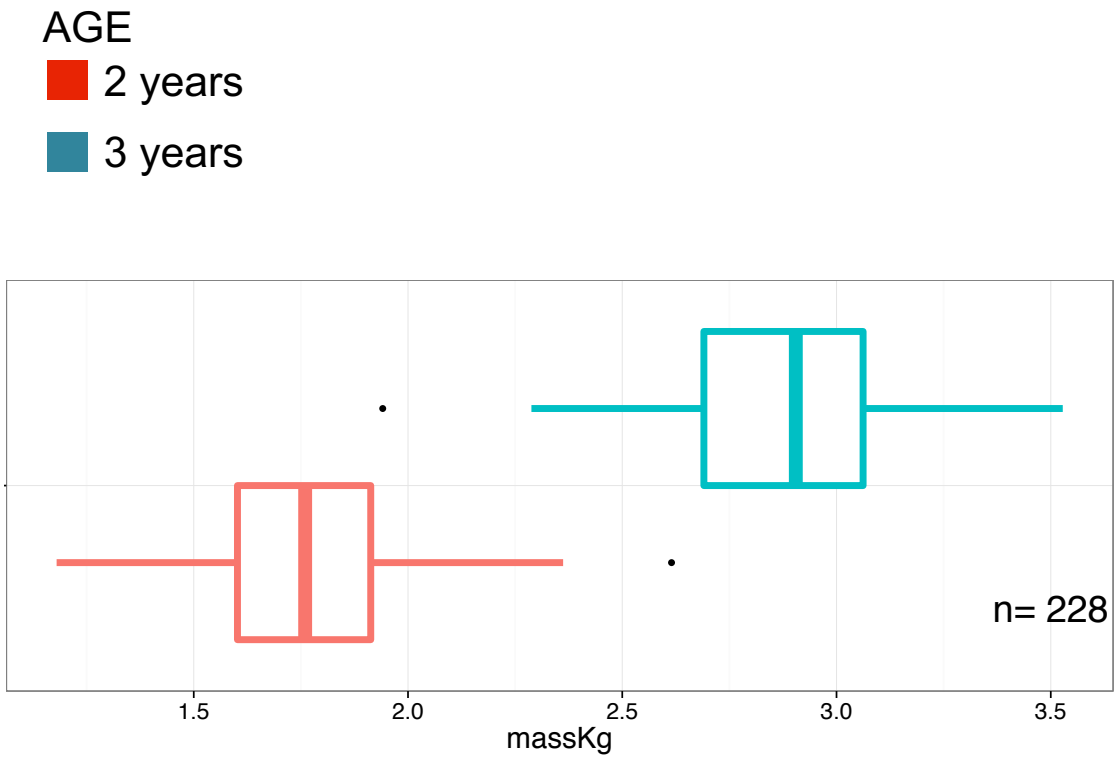
- Plotting **all data** points
- Getting a **sense of the data distribution**



boxplot

Plotting continuous data

Salmon body mass

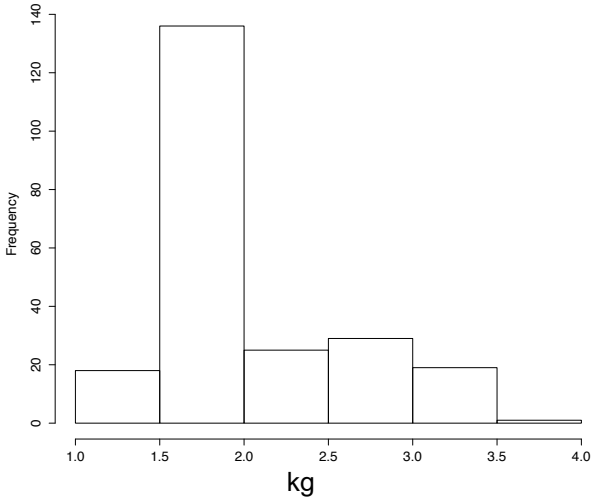


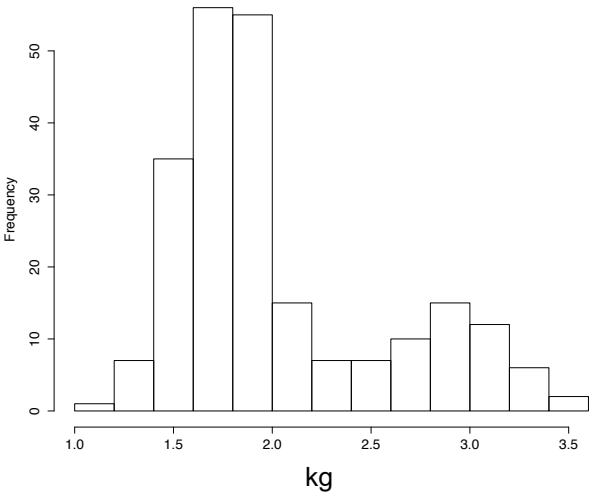
Additional **observations** can have huge **implications on the hypothesis**

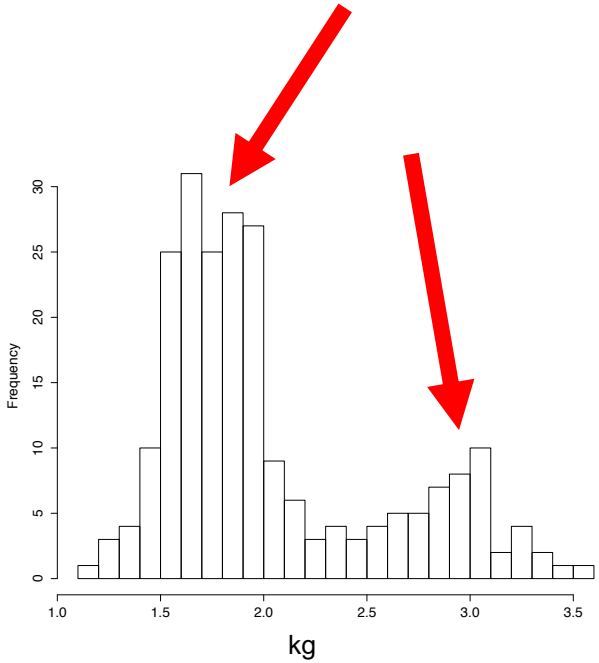
Correct **experimental design** help in **proper sampling**

Plotting continuous data with histograms

Salmon body mass

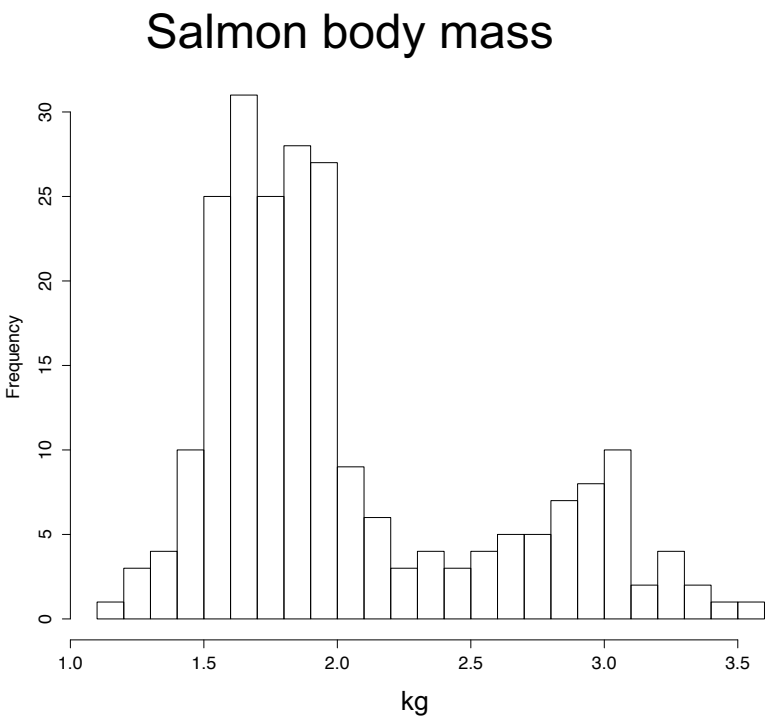


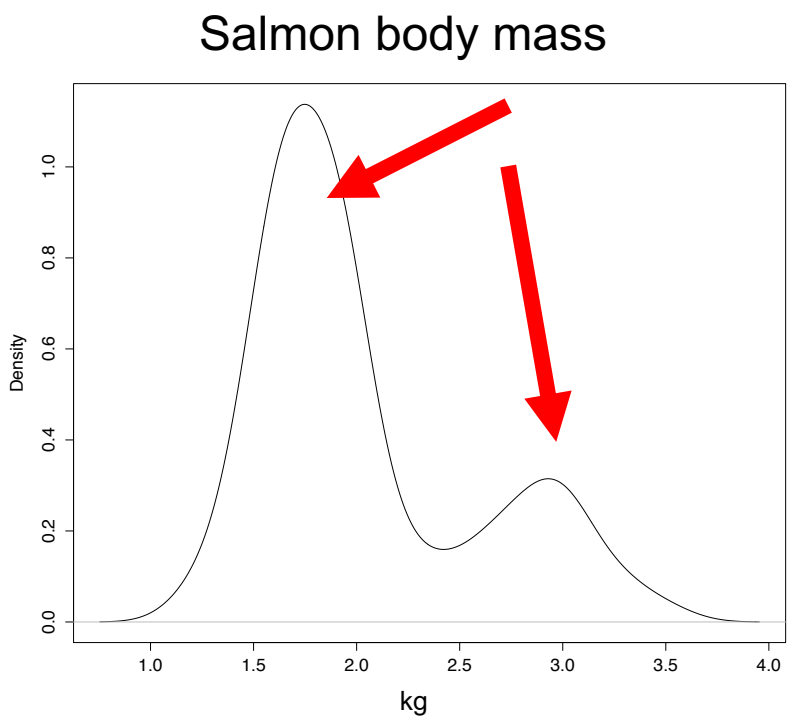




Be aware of the **bin width** (width of the bars)!!!

Plotting continuous data

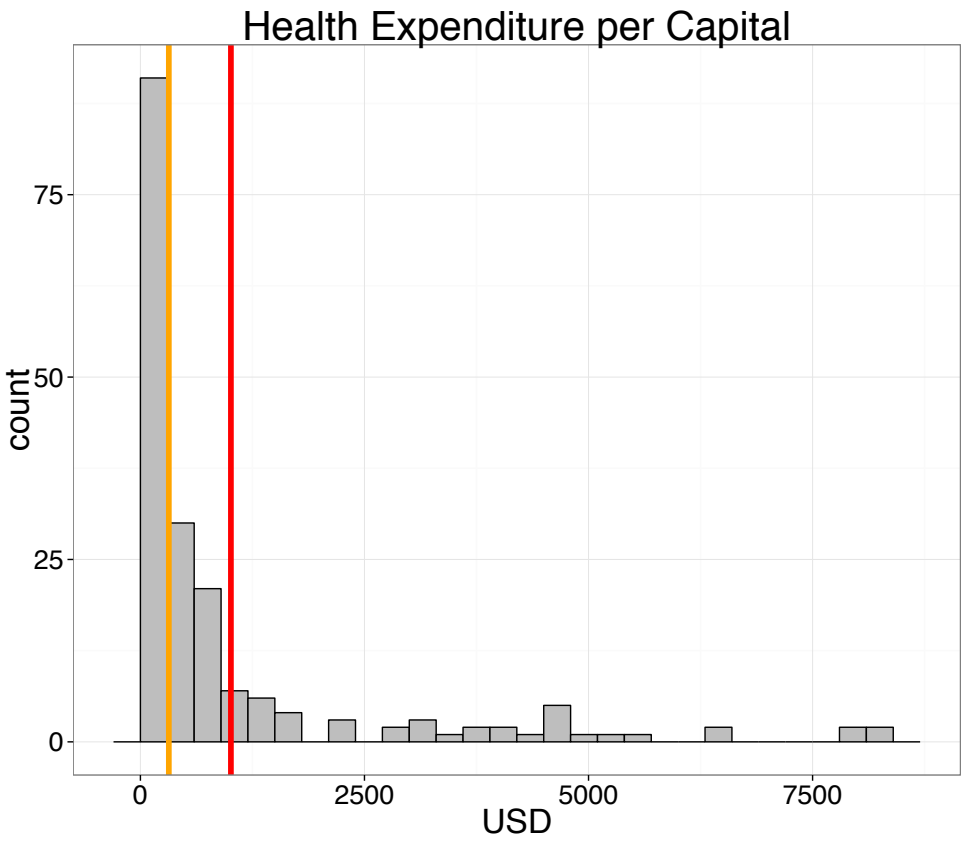




Data are smoothed automatically.
Density plots **blur discontinuities** in a distribution

non-Visual Data Description

Location and scatter



MEAN

sum of all observation/number of samples

MEDIAN

a number M such that 50% of the observations are less or equal to M, and 50% of the observations are higher or equal to M

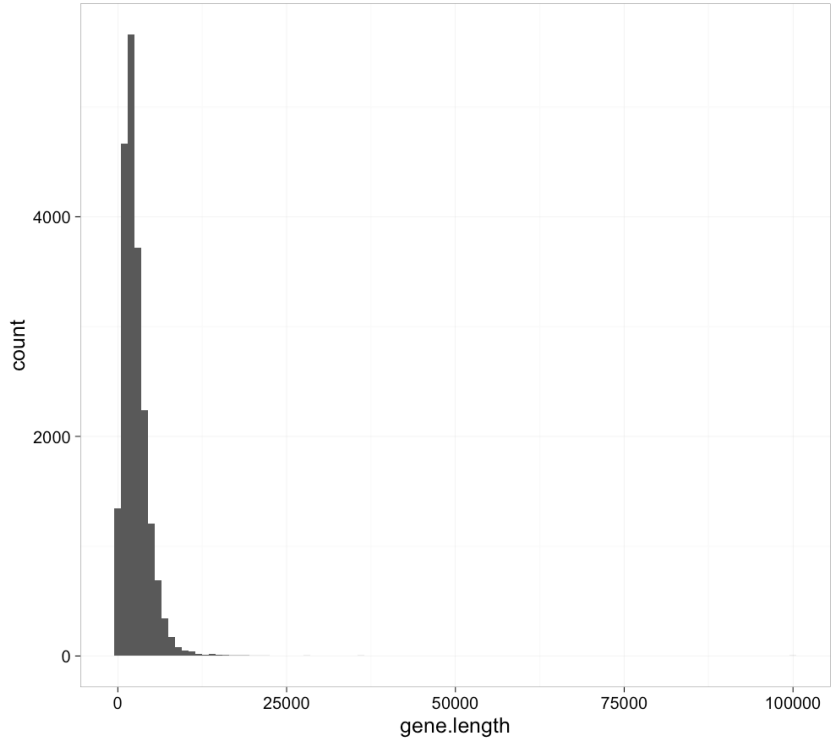
MEAN vs MEDIAN

MEDIAN should be preferred with:

- asymmetric distribution
- with extreme outliers

the **MEAN** is more precise on normal distribution

Location and scatter



MEAN

sum of all observation/number of samples

MEDIAN

a number M such that 50% of the observations are less or equal to M, and 50% of the observations are higher or equal to M

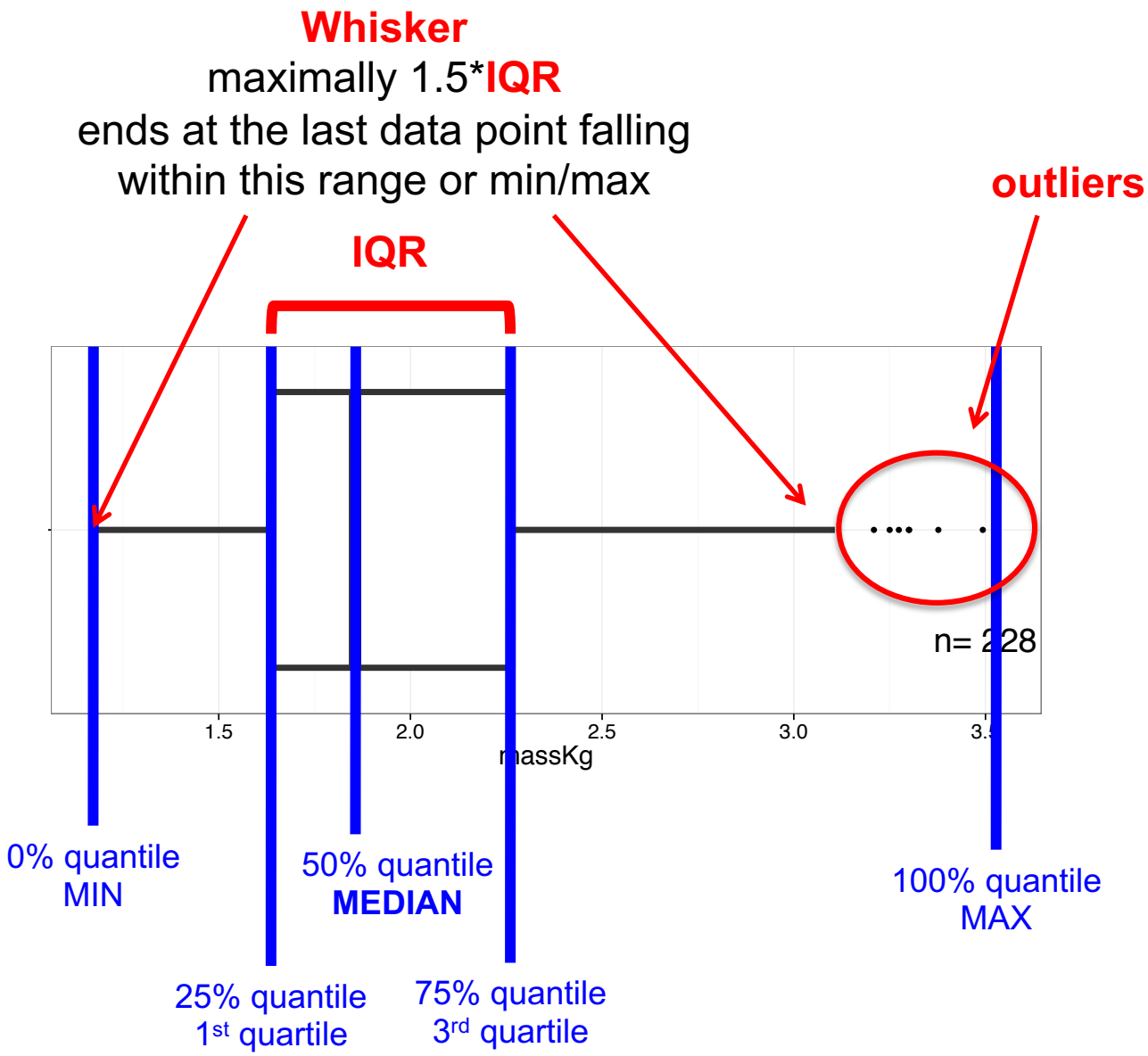
MEAN vs MEDIAN

MEDIAN should be preferred with:

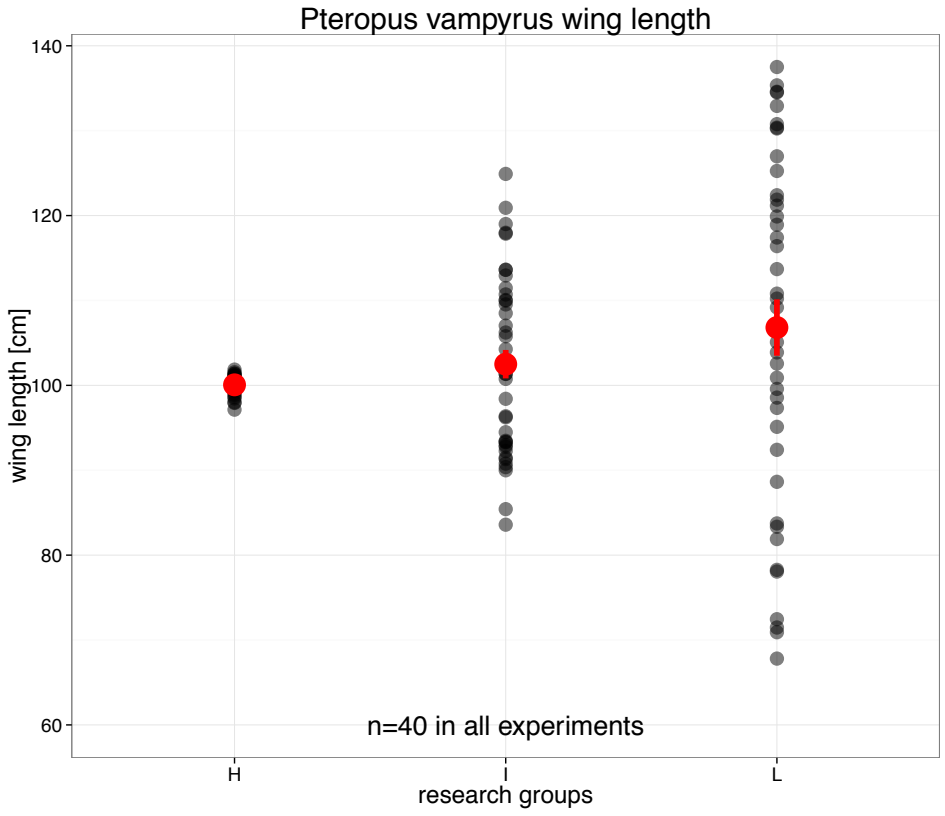
- asymmetric distribution
- with extreme outliers

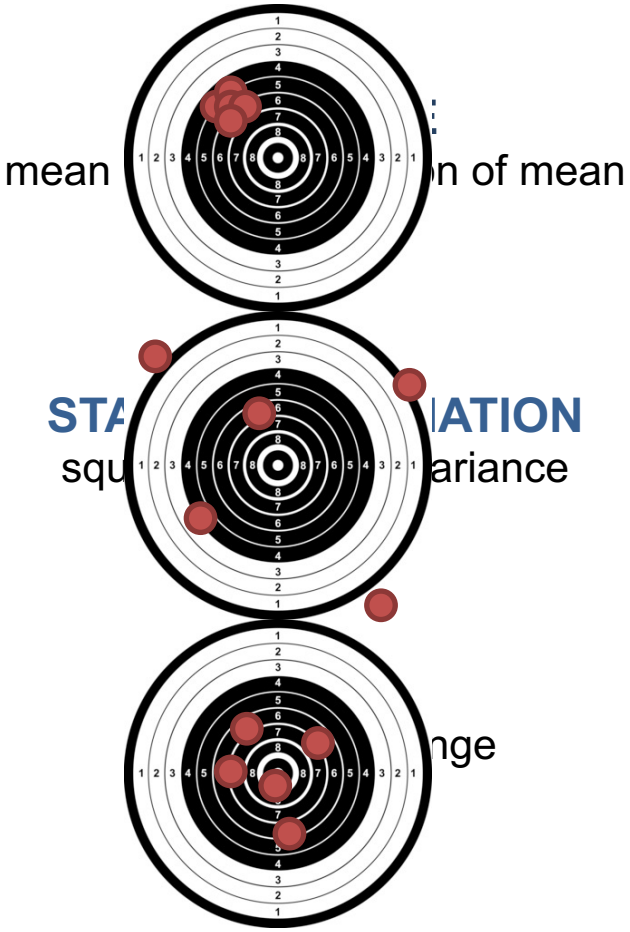
the **MEAN** is more precise on normal distribution

Quantiles – how to split distributions



Description of data scattering

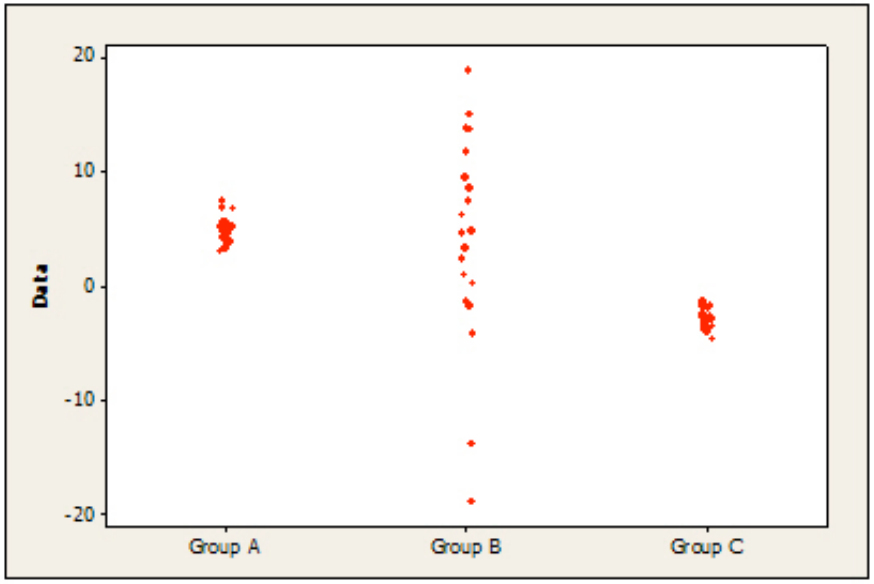




Comparing data-scatters

Homoscedasticity:
Two distribution have the **same scatter** (scatter comparable)

Heteroscedasticity:
Two distribution have **different scatter** (scatter not-comparable)



Group A and Group C exhibit homoscedasticity.

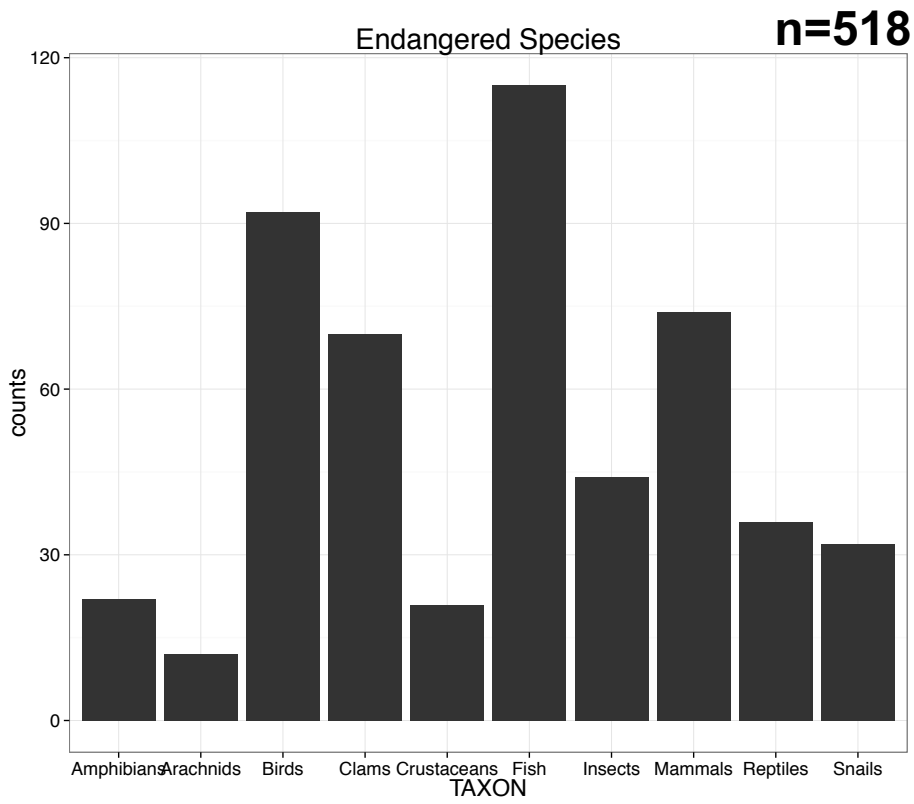
Group A and Group B exhibit heteroscedasticity (literally, "different scatter")

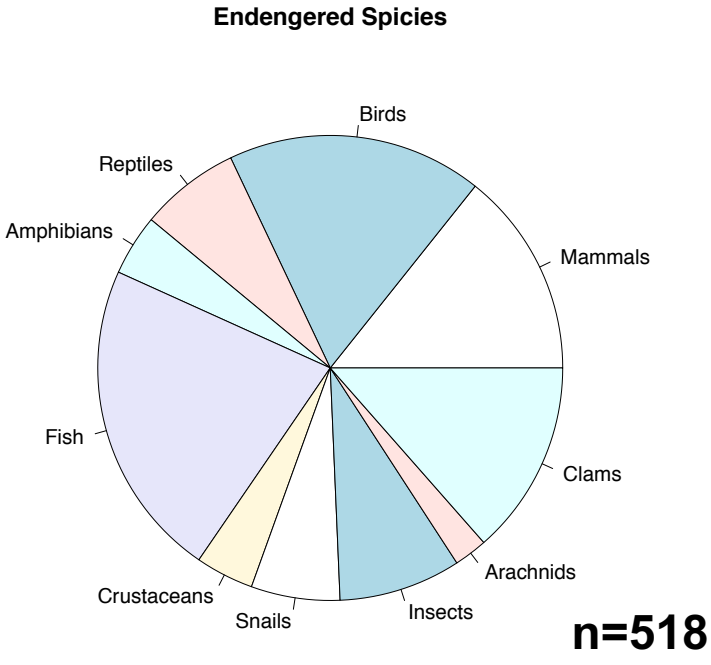
What about Groups B and C?

Categorical variables

	Mammals	Birds	Reptiles	Amphibians	Fish	Crustaceans	Snails	Insects	Arachnids	Clams	SUM
absolute frequency	74	92	36	22	115	21	32	44	12	70	518
relative frequency	14.3%	17.8%	6.9%	4.2%	22.2%	4.1%	6.2%	8.5%	2.3%	13.5%	100%

Categorical variables



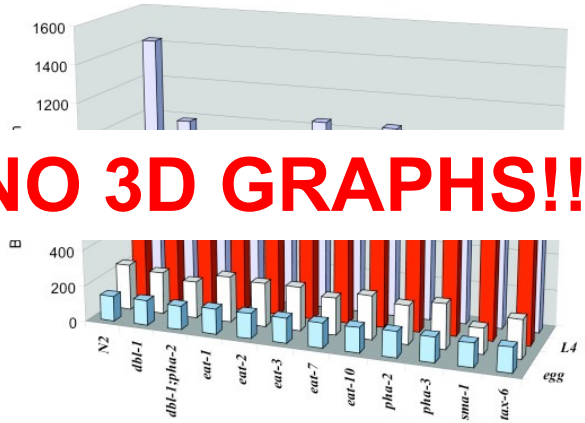


If you feel like doing a *pie chart*
...well...
STOP and use **TABLES** or **Bar charts** instead

Describing quantitative data

- Always report the **sample size**!!!!
- **Use numbers!** Median, Q1, Q3, scatter (for symmetric distribution mean and SD)
- **Use graphs!** Histogram, Boxplot, Density Plots
- **Tables** are one of the best way to summarize **categorical data**
- **Verbal** communication (***n***=*x*, ***IQR***=*y*-*z*)

Describing quantitative data



Body length in eggs, larvae and adults of various genotypes. Each bar shows the average of at least 25 measured individuals. See Table 3 for exact values and standard error of the mean.

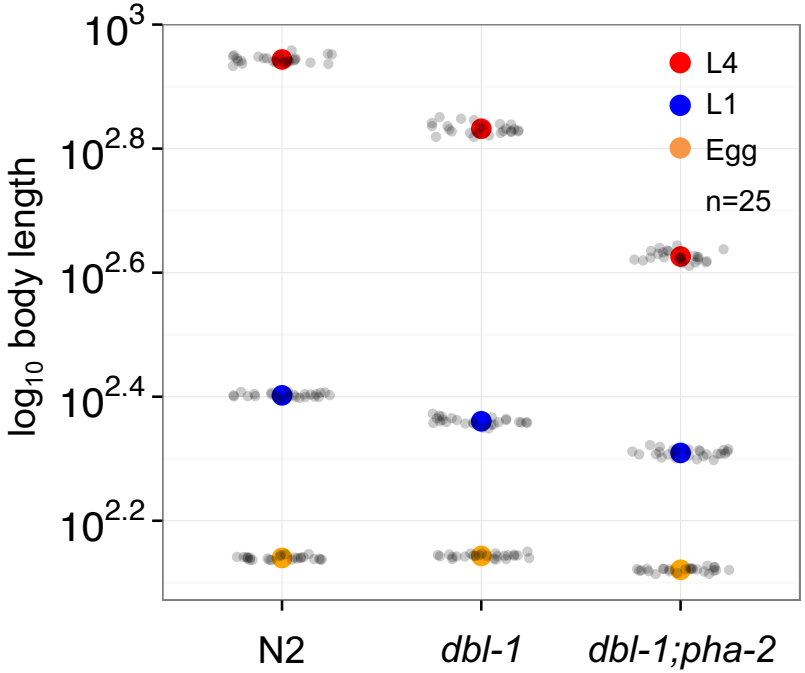
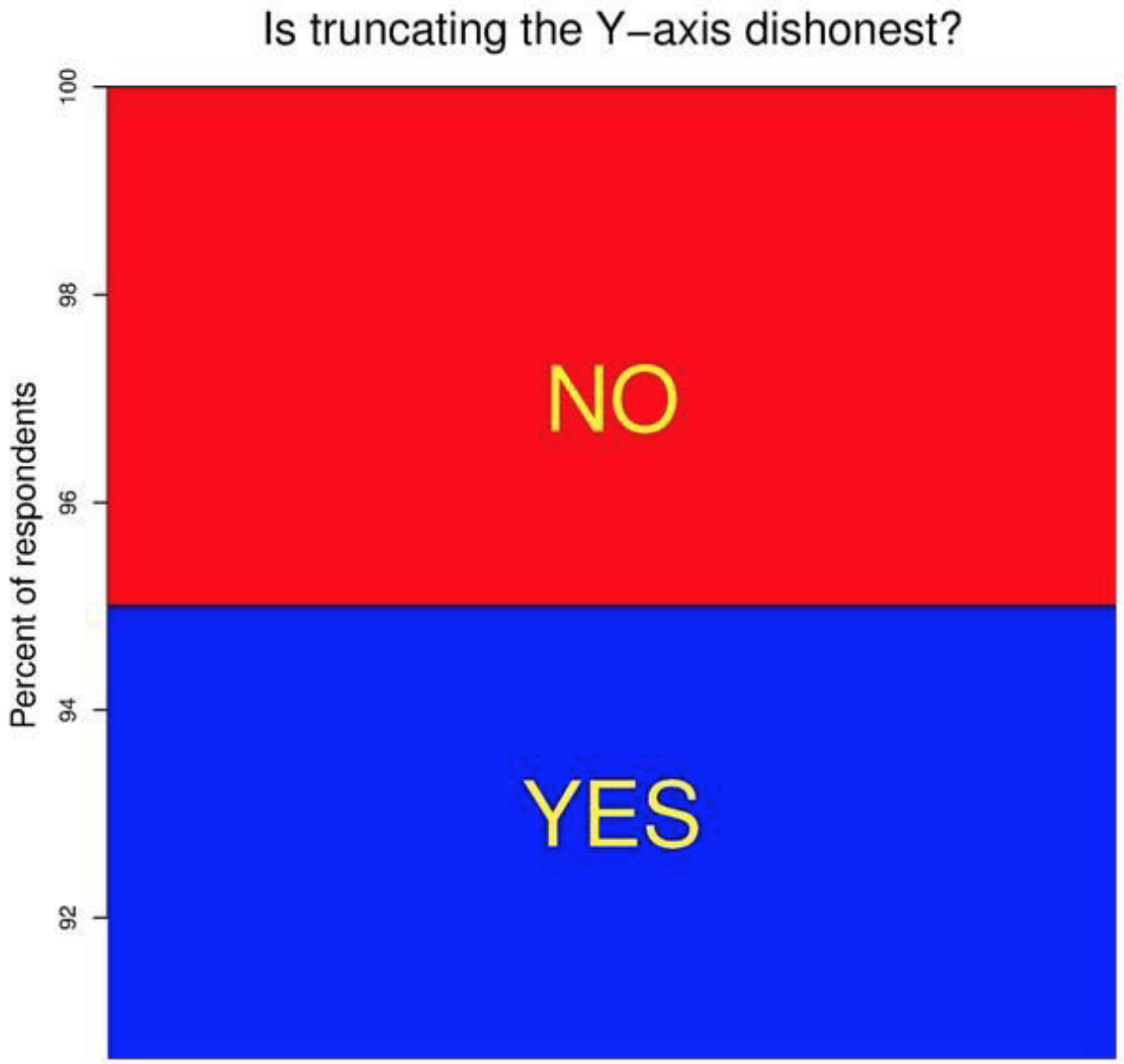


Table 3

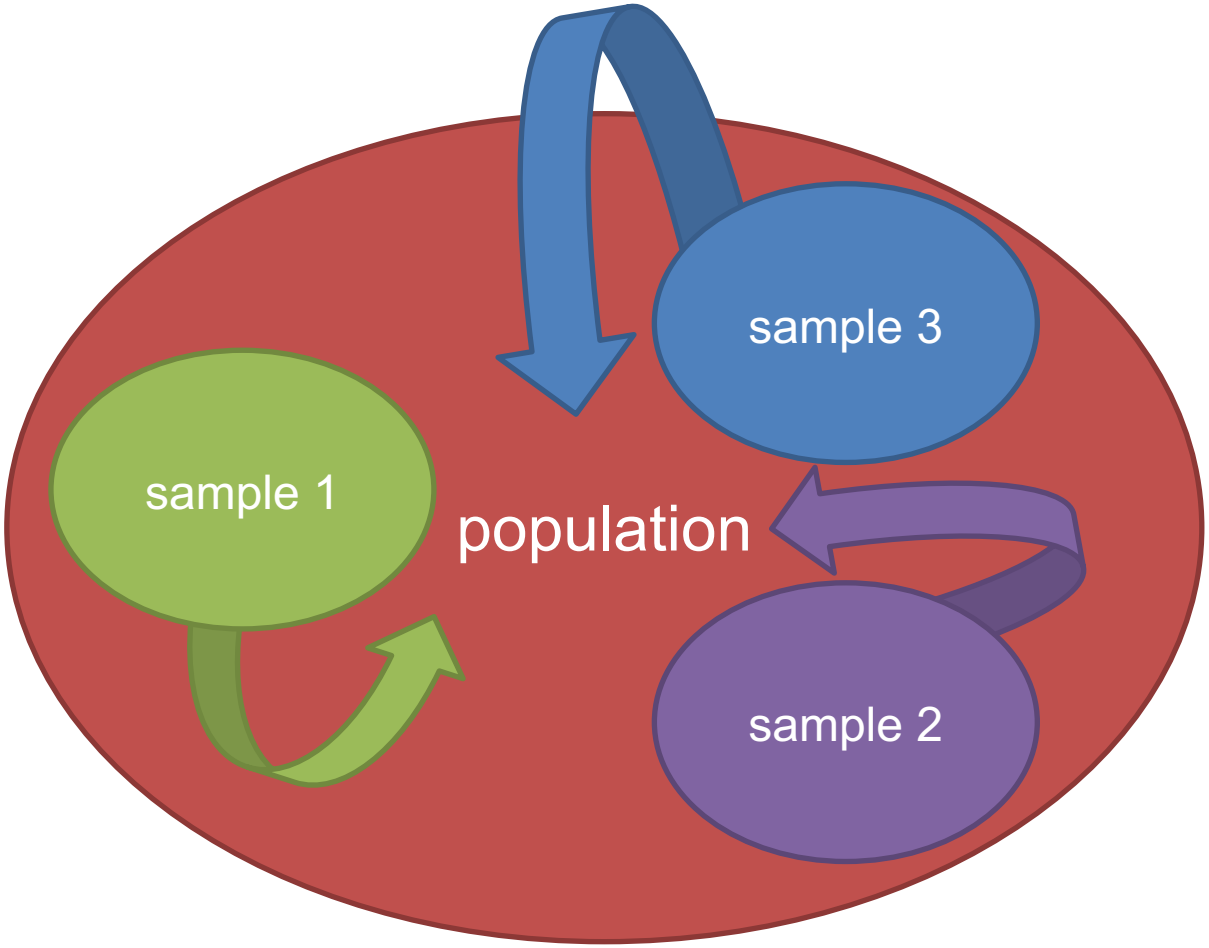
Study of eggs, L1 and L4 length. Eggs containing 3-fold stage embryos were measured. L1 larvae were measured within 1 hr after hatching and L4 larvae scored had an obvious white crescent surrounding the prospective vulva. At least 25 individuals were studied per genotype and experiment.

Genotype (in alphabetical order after N2)	**Egg circumference µm** (SEM)	**% circumference**	**L1 length µm** (SEM)	**% length**	**L4 length µm** (SEM)	**% length**
N2	138(± 0,2)	100	253(± 0,4)	100	878 (± 2,0)	100
dbl-1(nk3)	139(± 0,2)	101	230(± 0,7)	91	677 (± 2,6)	77
dbl-1;pha-2	132(± 0,2)	96	202(± 0,8)	80	423 (± 1,6)	48
eat-1(ad427)	140(± 0,2)	101	254(± 0,9)	100	696 (± 1,7)	79
eat-2(ad465)	139(± 0,2)	101	240(± 0,7)	95	762 (± 3,1)	87
eat-3(ad426)	138(± 0,3)	100	241(± 0,8)	95	734 (± 1,7)	84
eat-10(ad606)	139(± 0,2)	101	243(± 0,8)	96	747 (± 1,9)	85
egl-4(ad450)	138(± 0,2)	100	207(± 0,6)	82	676 (± 1,4)	77
pha-2(ad472)	141(± 0,2)	102	214(± 0,8)	85	688 (± 2,2)	78
pha-3(ad607)	138(± 0,2)	100	247(± 0,6)	98	654 (± 2,4)	75
sma-1(ru18)	132(± 0,2)	96	140(± 0,4)	55	607 (± 2,1)	69
tax-6(p675)	136(± 0,2)	99	213(± 0,7)	84	649 (± 1,4)	74

“I don’t mind lying, but I hate inaccuracy.” – S. Butler



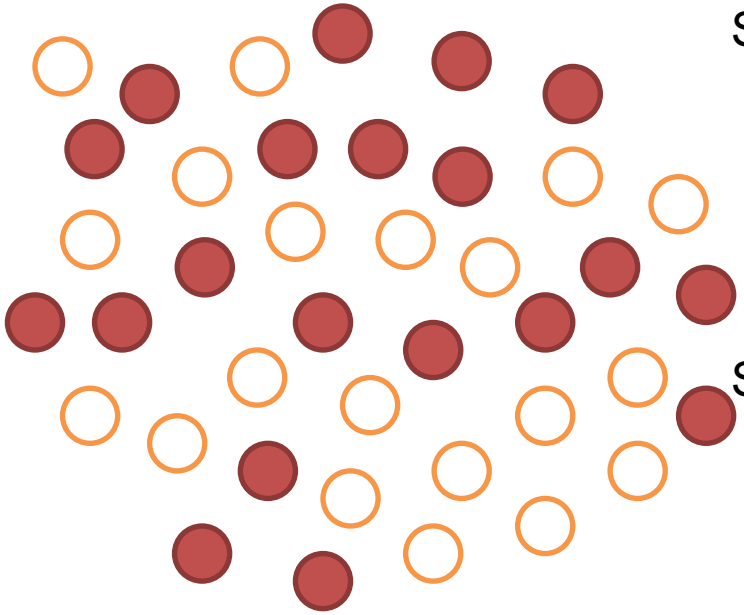
From the sample to the population



With **inference**/**induction** we refer to the process of describing the population starting from our samples.

From the sample to the population

We want to measure the incidence of mutation of the ***ade2*** gene in *S. cerevisiae*



Sample 1 = **1**:**5**

Sample 2 = **5**:**1**

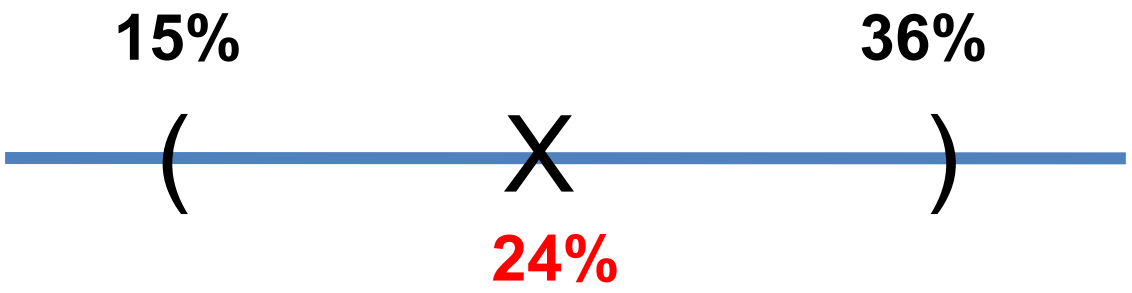
Sample 3 = **3**:**3**

1:**1**

TRUE VALUE = **50%**

Confidence Intervals

95%-Confidence Interval: An estimated interval which contains the "True Value" of a quantity with a probability of 95%



(1-α)-Confidence Interval: An estimated interval which contains the "True Value" of a quantity with a probability of (1-α).

α= error probability

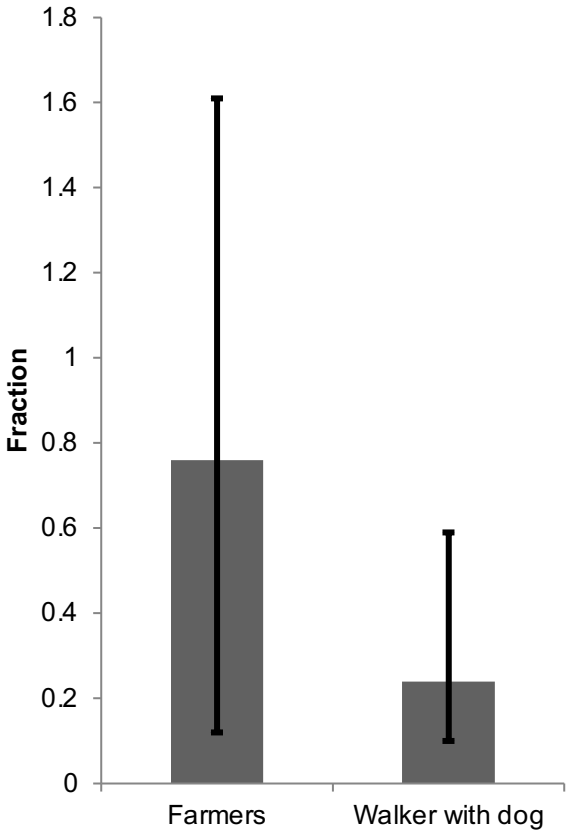
Proportional data

Cows are the most deadly large animals in UK. In the past 15 years, cows have killed 75 people. 57 of those were farmers, working closely to the animals. The remaining 18 were lonely walkers with dogs.
HSE-2015 - Independent

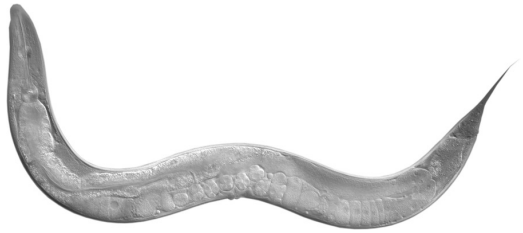
How can we represent the data?
This is a **binomial test**, and we can represent it with the **95% CI**

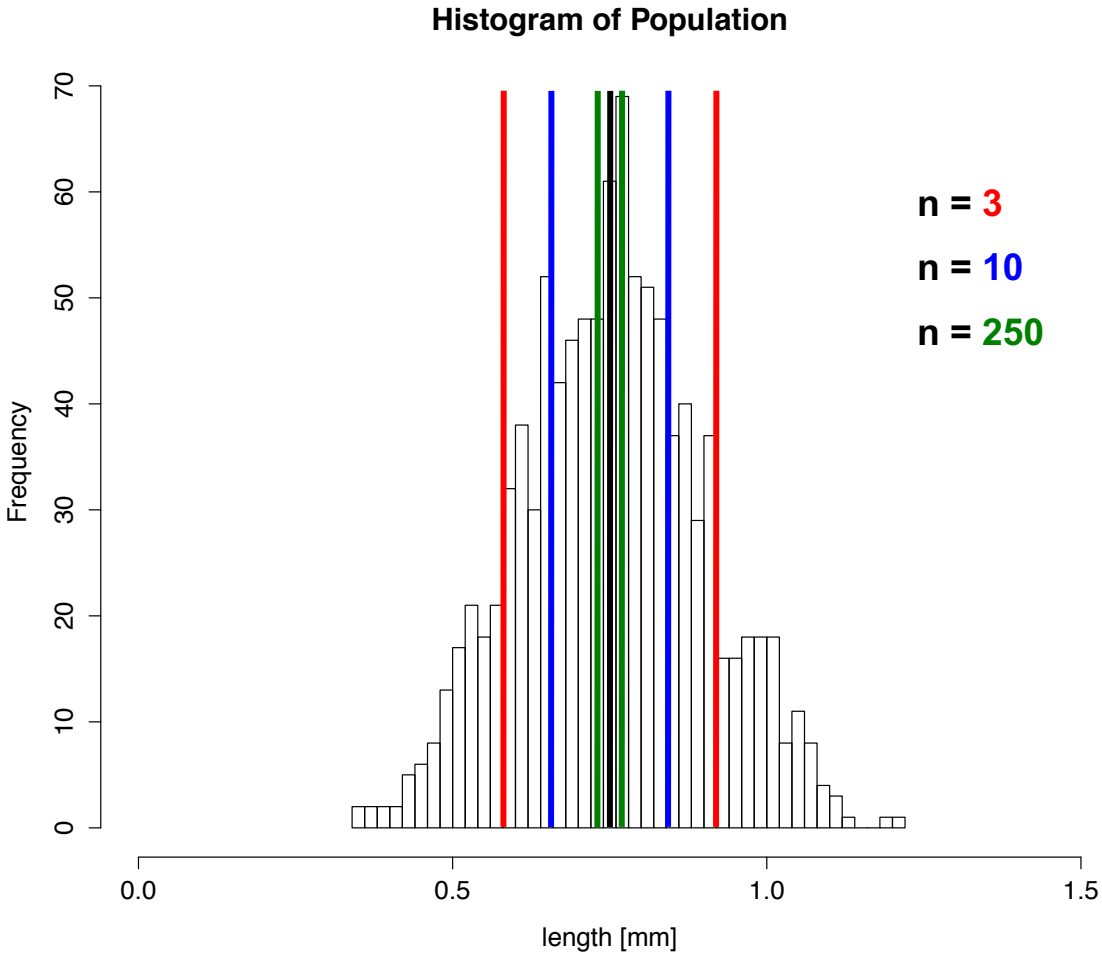
http://statpages.org/confint.html

What is the assumption we have to make?
Selection of sample was random.



Sample size and confidence interval

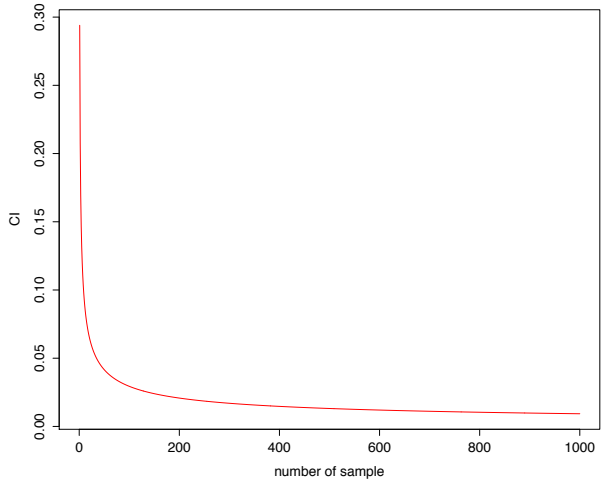




n = 3

n = 10

n = 250



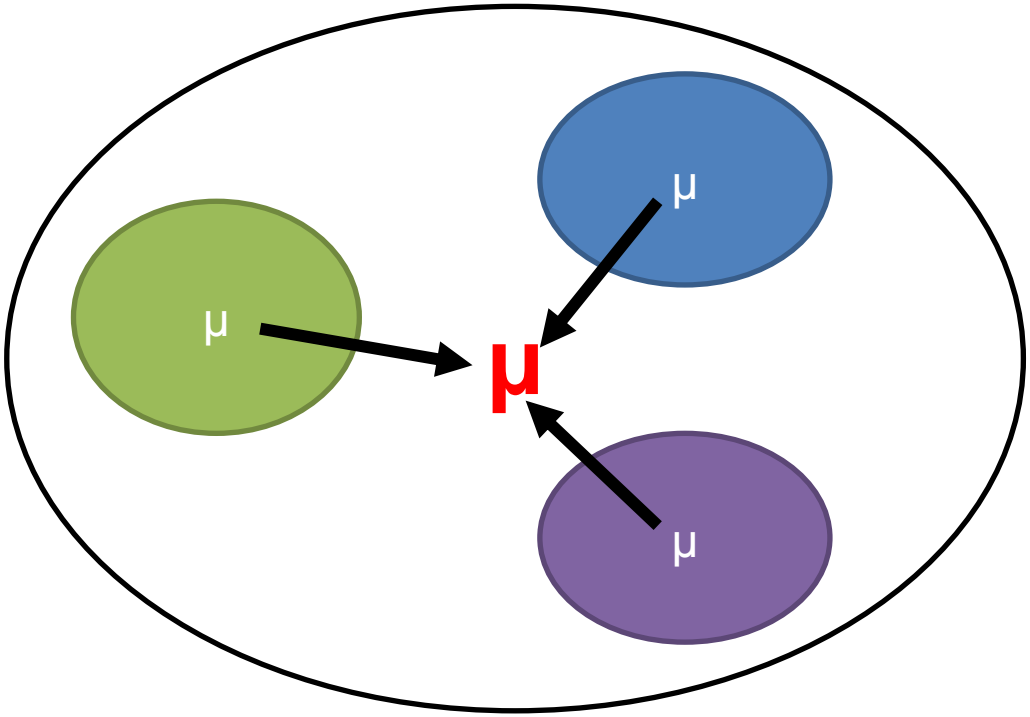
Standard Error of the Mean (SEM)

Standard Error of the Mean (**SEM**) is the standard deviation (**SD**) of the **sample mean** estimate of a **population mean**

SEM = SD / square root (**n**)

The **smaller the SEM** the **closer** we are to the **true population**

The **larger the SEM** the **farer** we are from the **true population**



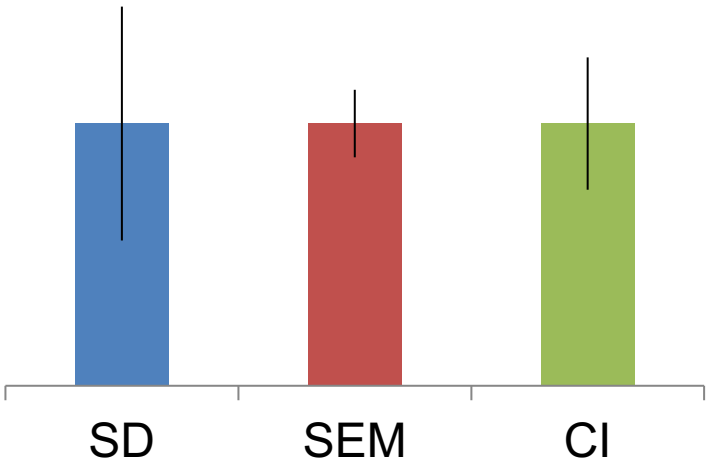
Standard Error of the Mean (SEM)

When we want to describe a certain population we need to use the **SEM.**

Some people prefer the **SEM** to the **SD** because of the smaller size (???).

If two **SEM overlap**, the two experiments are surely **not different** (statistically significant)

If two **SEM DO NOT overlap**, it is **NOT certain** that the two experiments are **significantly different**



SD can be used when we **describe the distribution** without aiming to describe the entire population.
e.g. **Technical replicates!!**

The Journal's "RULE"

- **ALWAYS report the N** in the graph or in your figure legends! (e.g. sample size or the number of independently performed experiments)

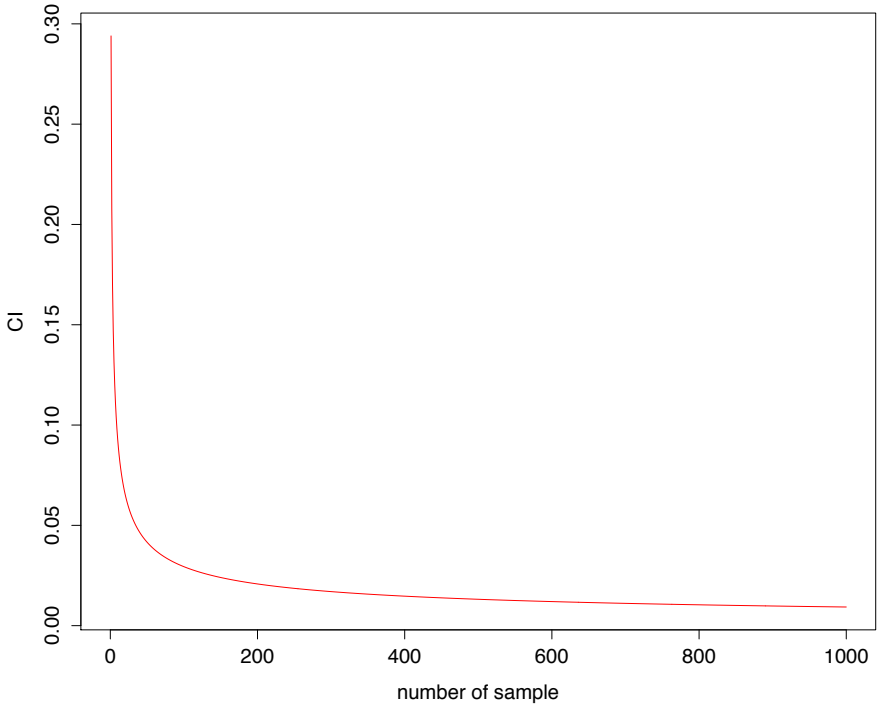
- Error bars and statistics must be reported **ONLY** for **independently repeated experiments**. **NEVER** for **technical replicates**. In some cases, you can report a "*representative experiment*", which doesn't have error bars or p-values, because n=1

- In biology we usually **compare experiments to controls**, it is appropriate to report **inferential error** bars, such as **SEM** or **CI** rather than SD.

- Reporting **all the individual data points** is usually better, especially when **N is small.**

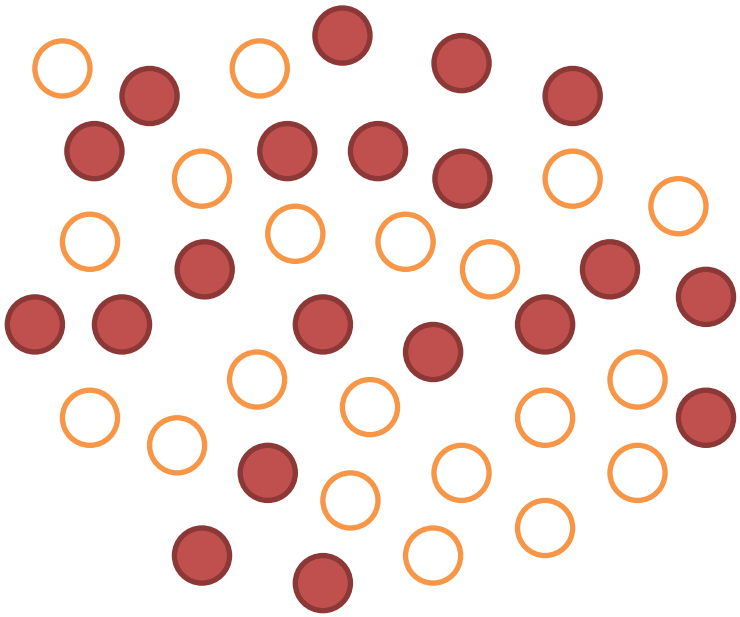
From the **Sample** to the **Population**

Large sample number**!**



Does a **limit** exist?

Random sampling



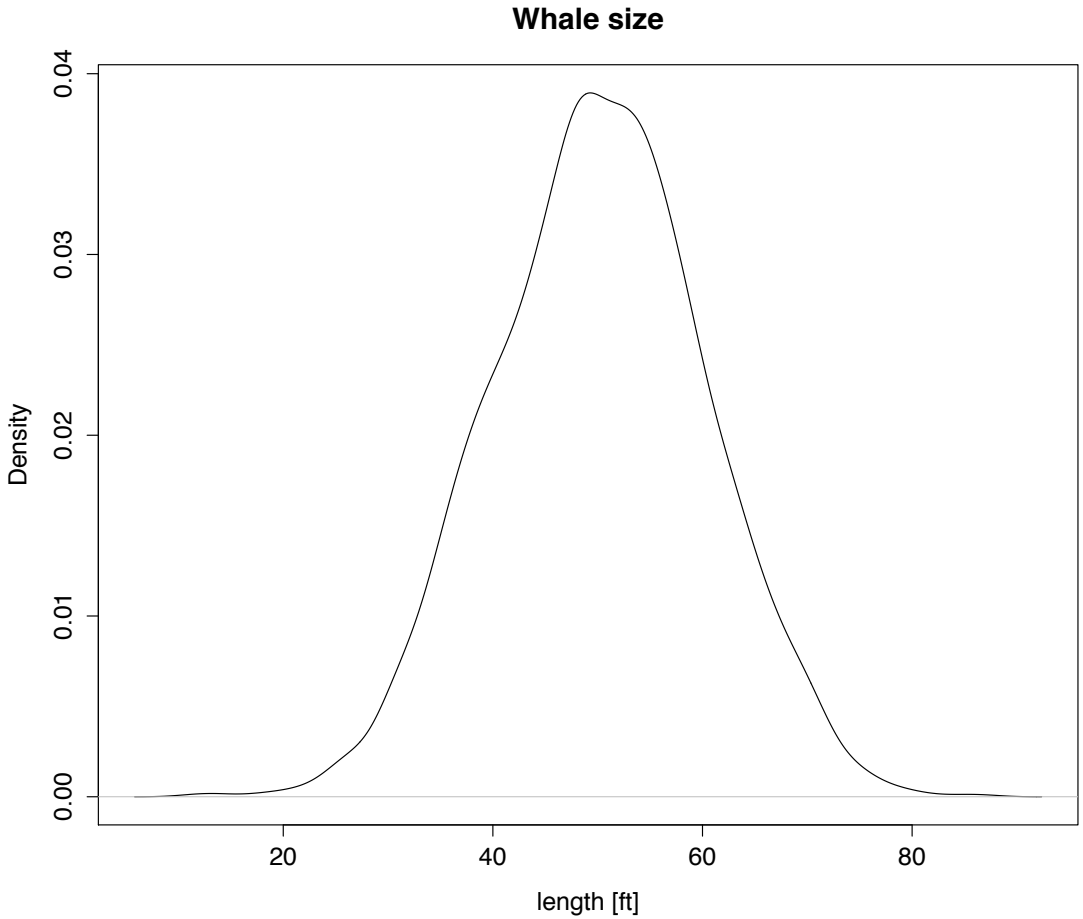
How to be **random** in the sampling method?
Is it possible?

Test

Measuring whales in the Atlantic Ocean

Training-phase:

measuring all the whales that we encounter in the ocean

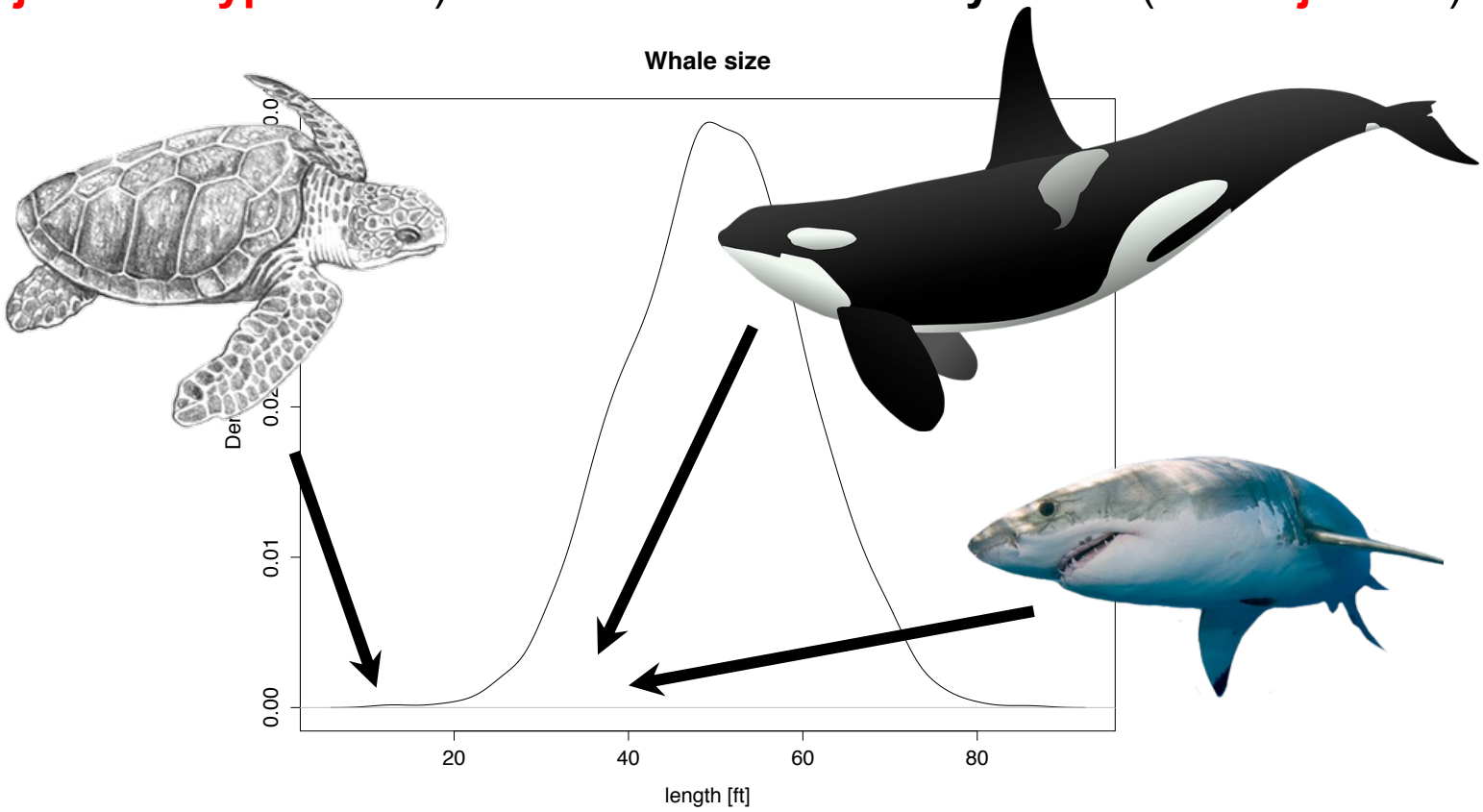


Measuring whales in the Atlantic Ocean

Test-phase:

For any unknown animal, test the hypothesis that it is a whale.

The only thing we know about whales is their length. If the length of the animal is **out of the bounds**, then we call the animal a non-whale (**we reject the hypothesis**). Otherwise...**we can't say much** (**non-rejection**).



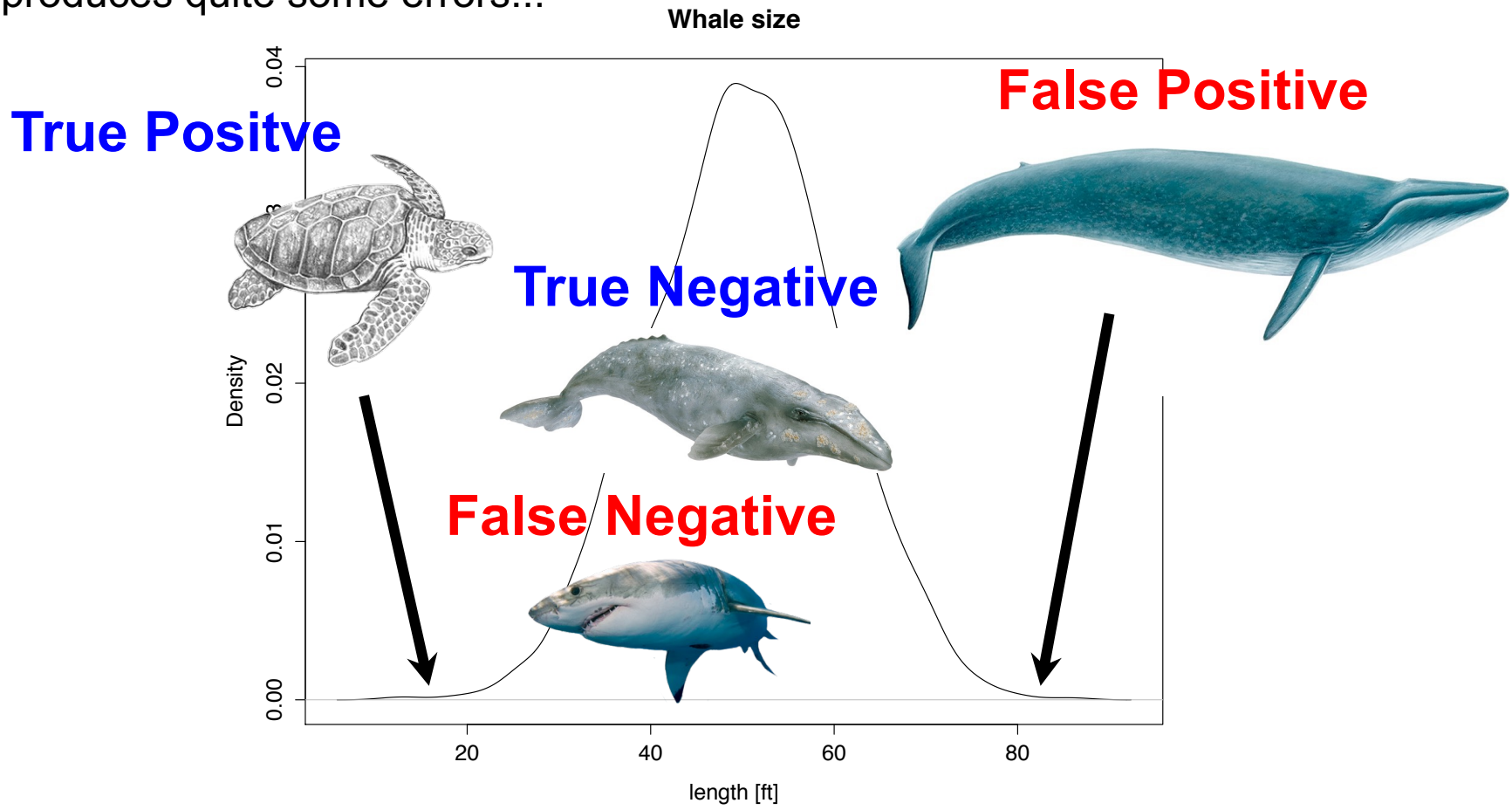
Measuring whales in the Atlantic Ocean

Advantage of the method:

we don't need to know much about whales...

Disadvantage:

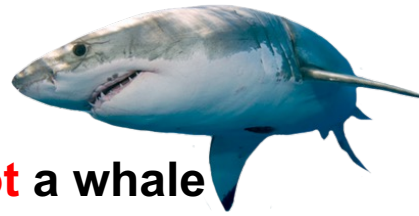
It produces quite some errors...



Measuring whales
in the Atlantic Ocean

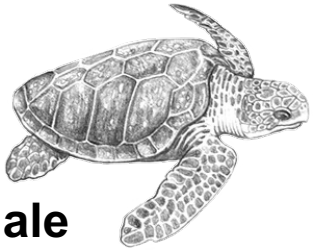
True Negative

it is a **measurement that doesn't reject** the hypothesis and **it is a whale**



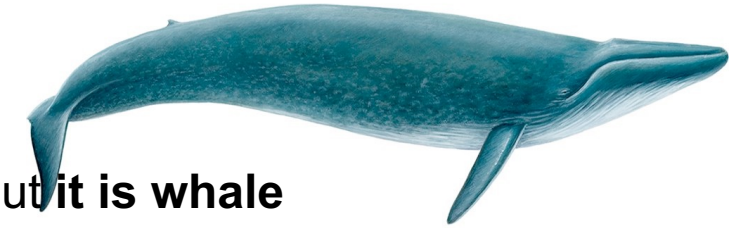
False Negative

It is a **measurement that doesn't reject** the hypothesis but it **is not a whale**



True Positive

It is a **measurement that rejects** the hypothesis and **it is not whale**



False Positive

It is a **measurement that rejects** the hypothesis but **it is whale**

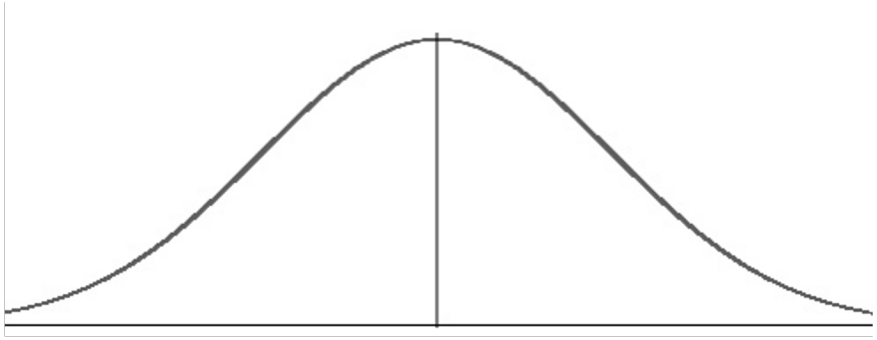
Measuring whales in the Atlantic Ocean

State a **null hypothesis H_0**

e.g. *“there is no difference, nothing changes...it is a whale”*

Choose an appropriate **test statistic**

the data-derived quantity that leads to the decision for the rejection.
Implicitly it determines the null distribution (the distribution of the test statistic under the null hypothesis).



These steps belong to the **EXPERIMENT DESIGN** phase!!!

You don’t make up your mind according to the data you were able to collect...

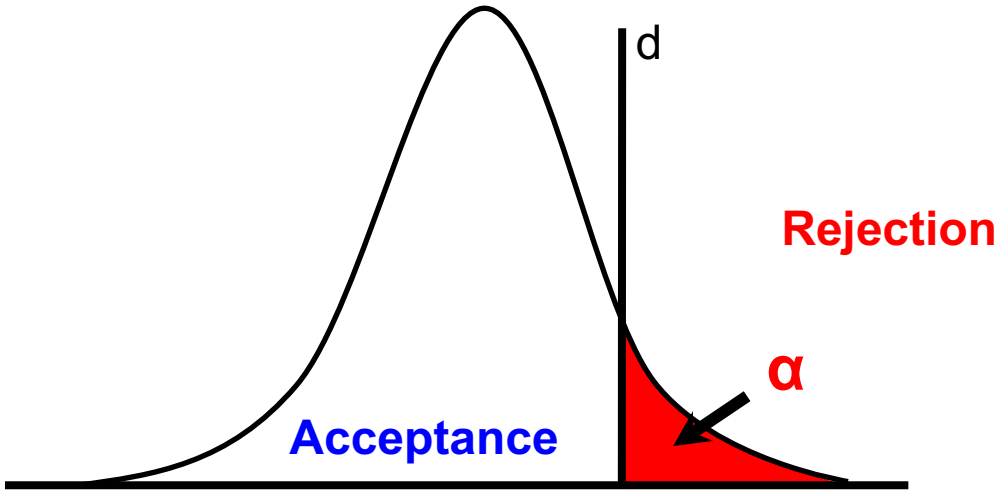
Measuring whales in the Atlantic Ocean

State an **alternative hypothesis H_A**

e.g. *"there is difference...it is not a whale"*

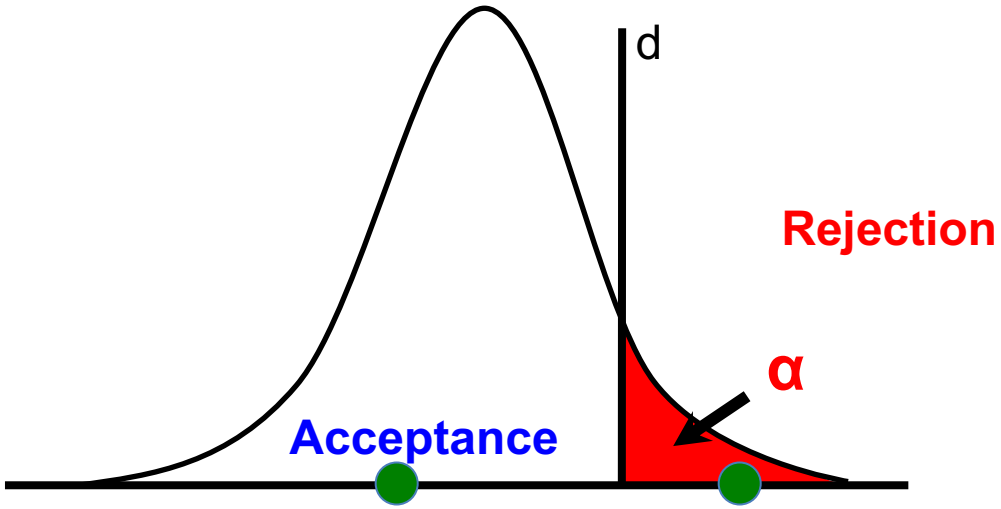
Determine **decision boundary** for the rejection.

In other words, define a **significance level α**, i.e. the fraction of **false positive** calls you are willing to accept.



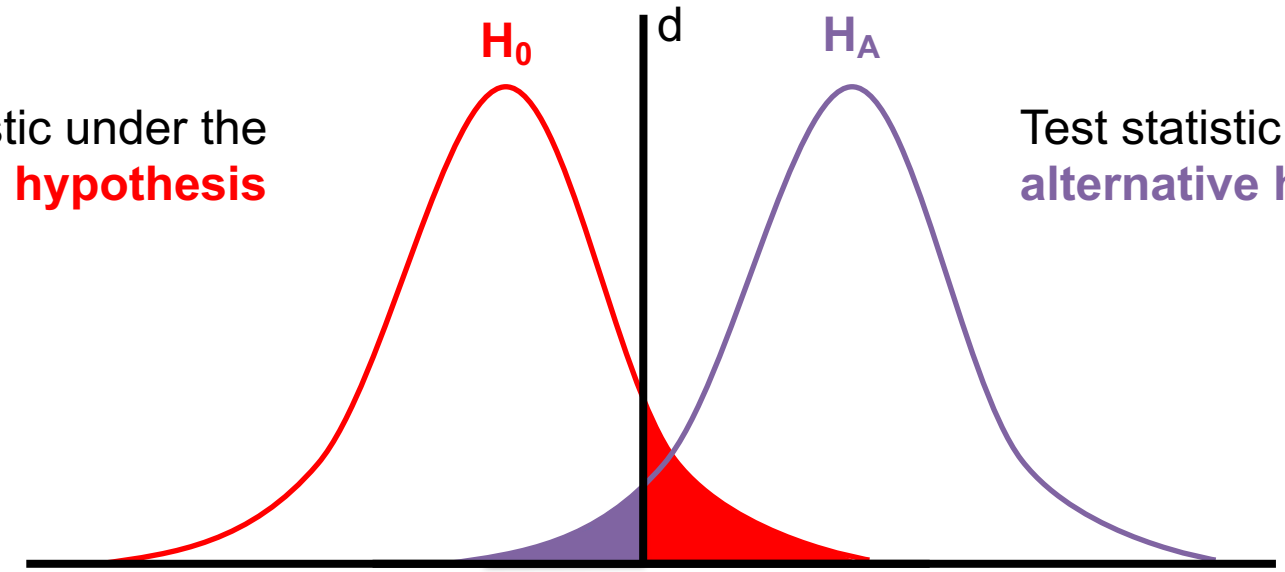
Measuring whales in the Atlantic Ocean

Calculate the actual **value** of the test and see whether you will reject or not the hypothesis according to the **pre-specified (!!!)** decision boundary.



Good Test Statistics

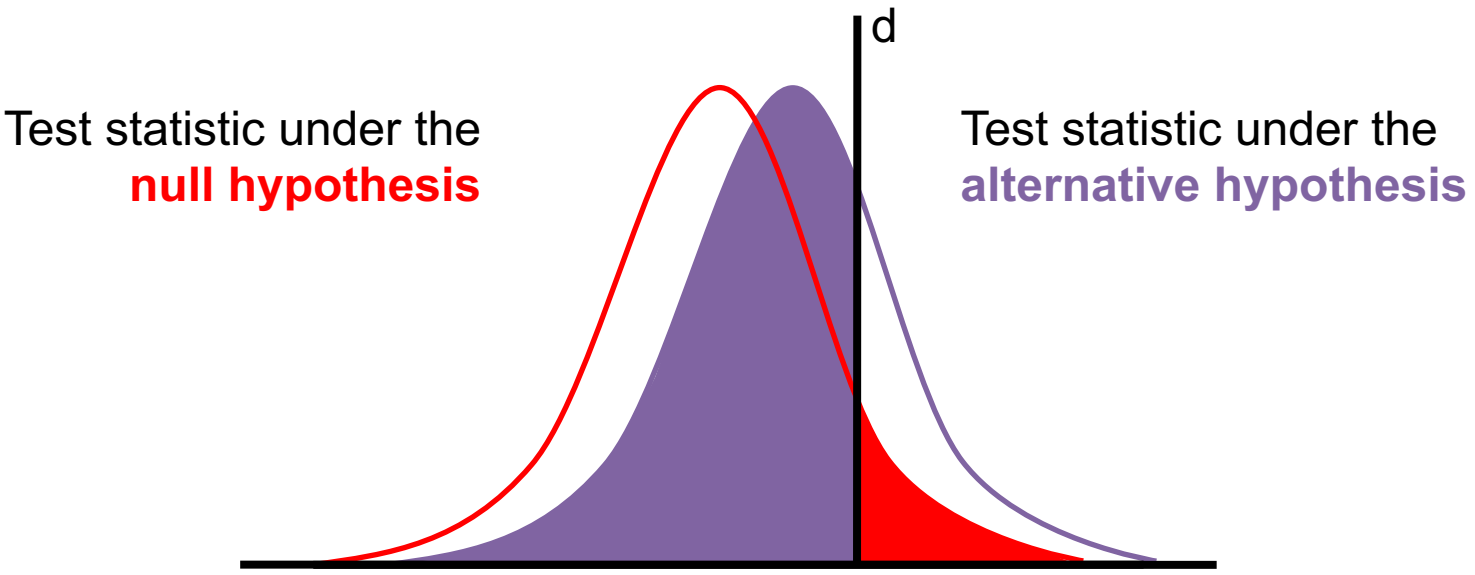
Test statistic under the **null hypothesis**



Test statistic under the **alternative hypothesis**

	Accept H_0	**Reject H_0**
H_0 TRUE	correct decision	**Type I Error "False Positive"**
H_A TRUE	**Type II Error "False Negative"**	correct decision

Bad Test Statistics

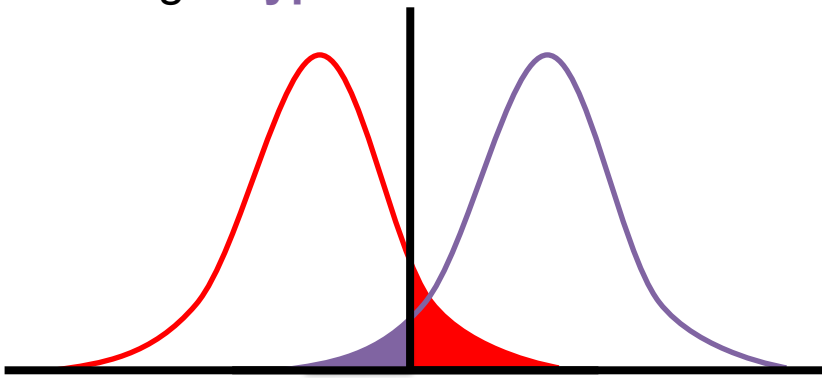


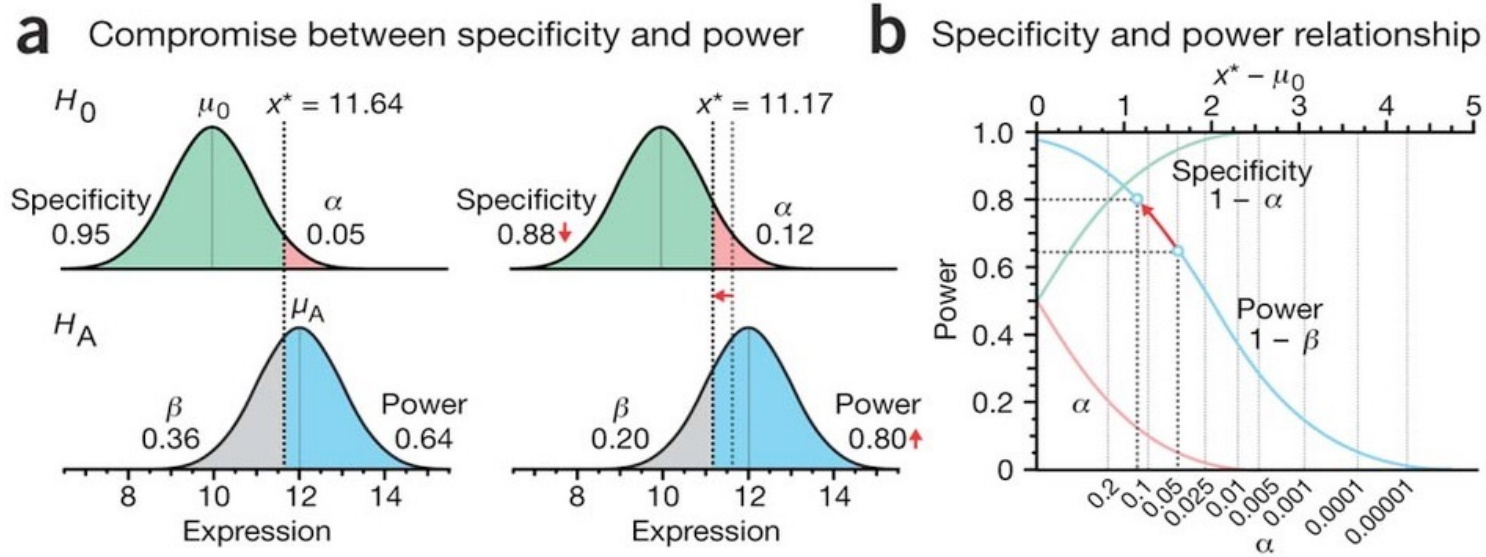
	Accept null hypothesis	Reject null hypothesis
H_0 TRUE	correct decision	**Type I Error "False Positive"**
H_A TRUE	**Type II Error "False Negative"**	correct decision

Statistical Power

Probability that the test **will reject the null hypothesis** when the **alternative hypothesis is true.** It is the probability of **not** committing a **Type II Error**

The chances of committing a **Type II Error** decrease with the increase of the **statistical power**. The probability of a **Type II Error** is referred as the **false negative rate (β)**. Thus, the statistical power is defined as **1-β**. Interestingly, **1-α** defines the **specificity** of the test.



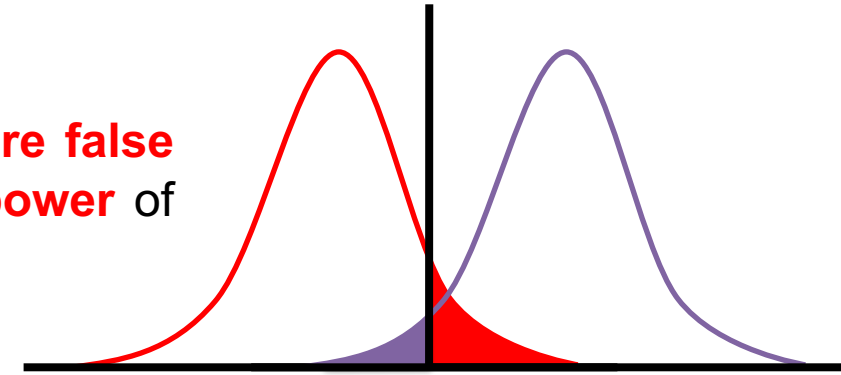


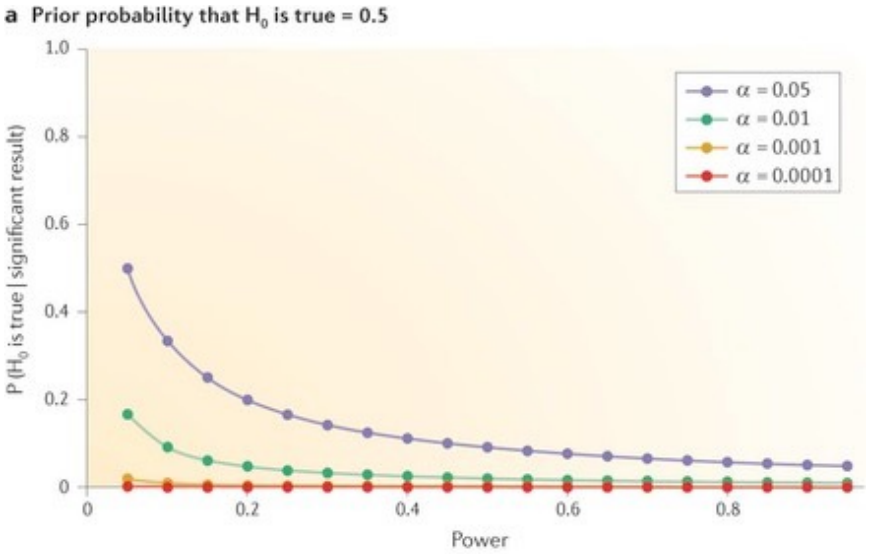
M. Krzywinski & N. Altman , Nature 2013

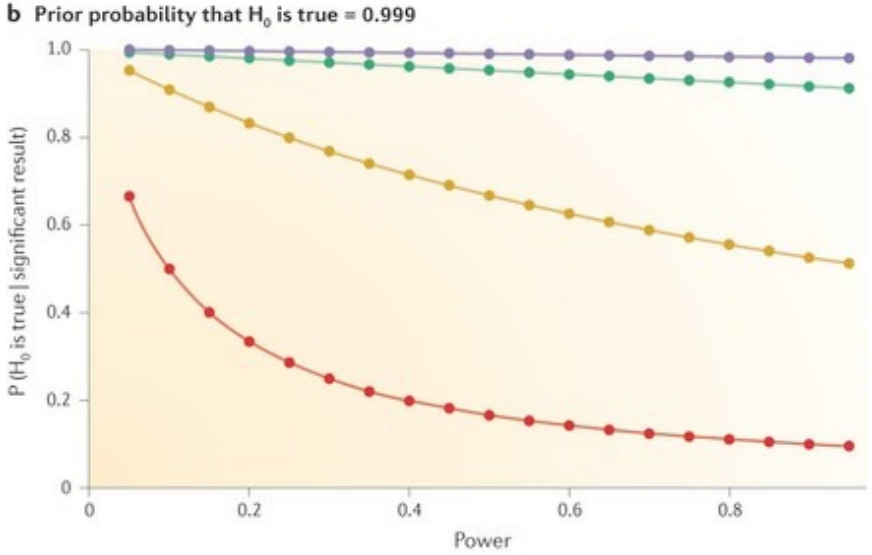
Statistical Power

It is completely **wrong** to assume that **Type I Error** (or False Positive) rates are **independent of the statistical power**.

In fact, **most significant published results are false positives** also thanks to the **low statistical power** of the test.







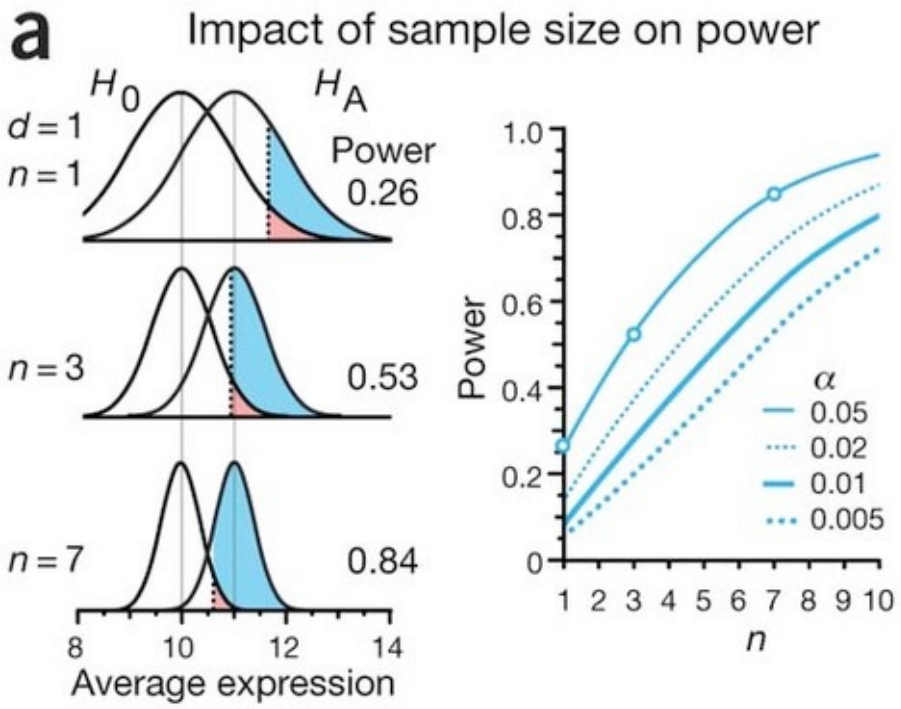
P.C. Sham & S.M. Purcell , Nature 2014

Nature Reviews | Genetics

Statistical Power

How can we be **specific (1-α)** while having **high statistical power(1-β)**?

We need to **increase the N** of our experiments. As we have seen previously, higher the N narrower the distribution.

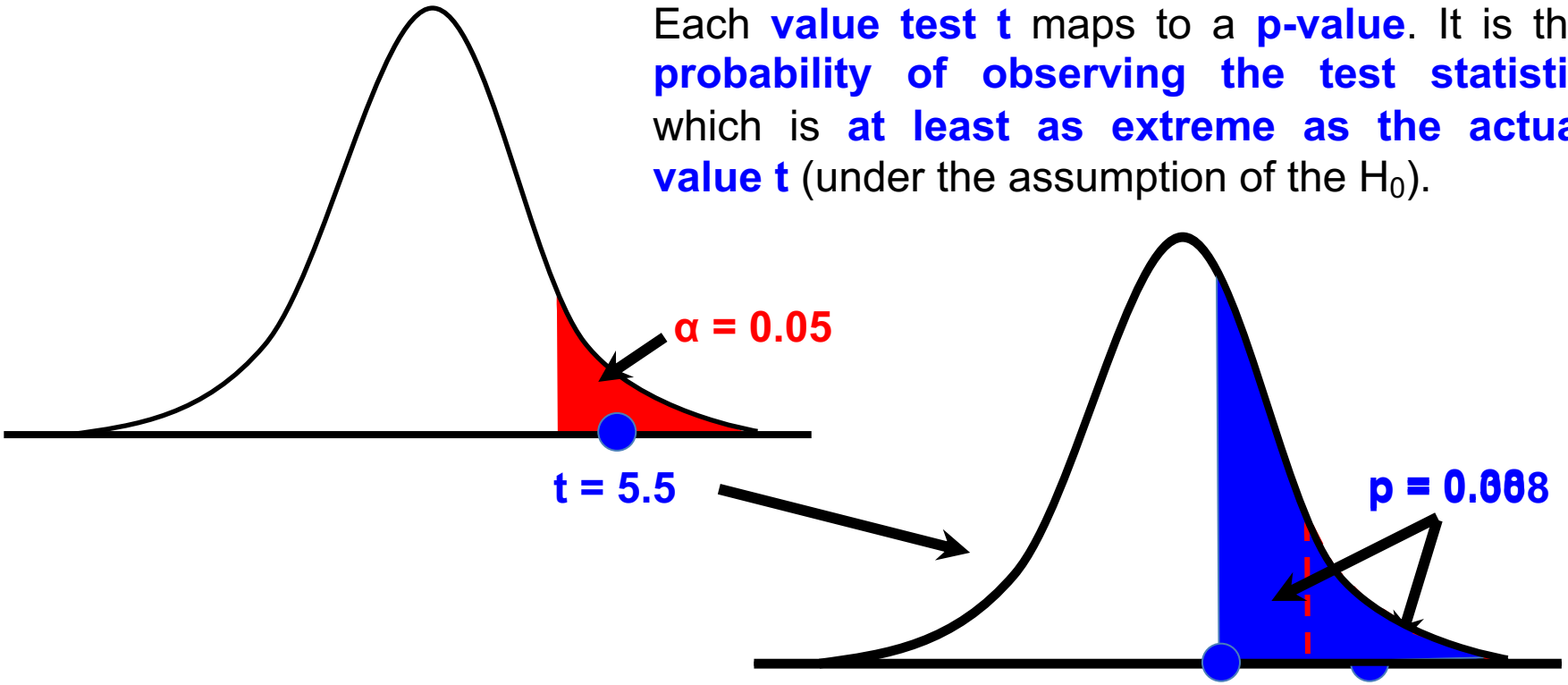


M. Krzywinski & N. Altman , Nature 2013

The p-value

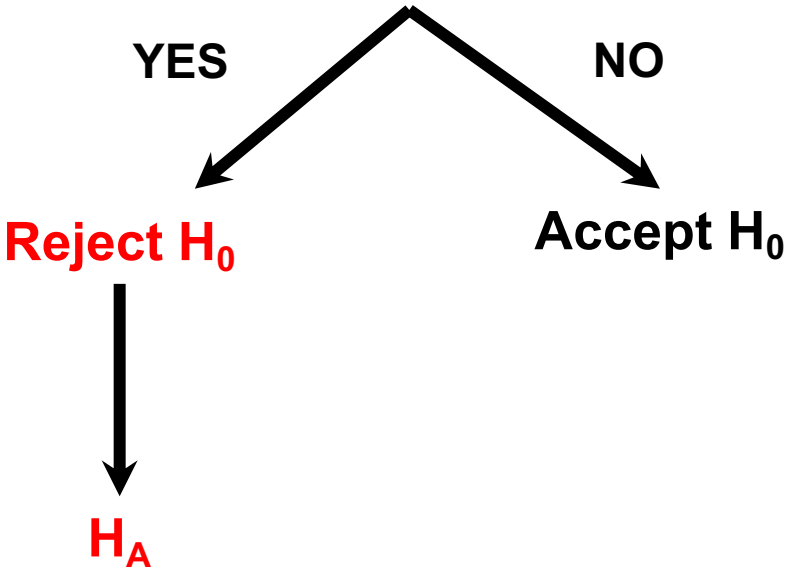
"Remember, a ***p-value* is not a measure of how right you are or how important a difference is**. Instead, think of it as a **measure of surprise** [emphasis added]. If you assume your medication is ineffective and there is no reason other than luck for the two groups to differ, then **the smaller the p value, the more surprising and lucky your results are** – or your assumption is wrong, and the medication truly works." – A. Reinhart, "*Statistics done wrong".*

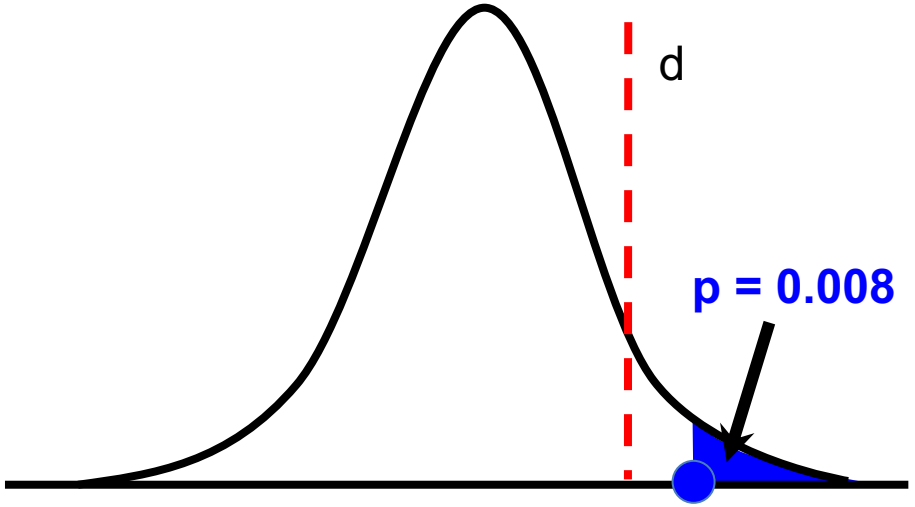
Each **value test t** maps to a **p-value**. It is the **probability of observing the test statistic** which is **at least as extreme as the actual value t** (under the assumption of the H_0).

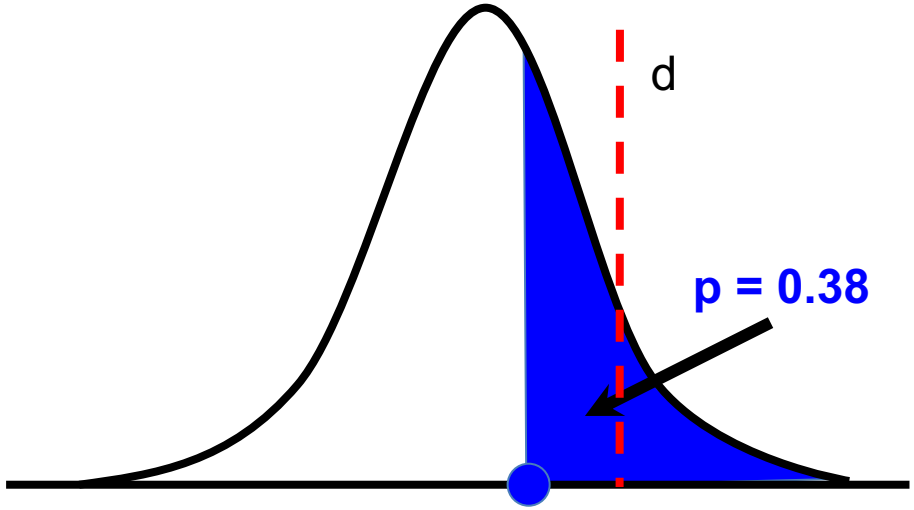


The p-value

Is the **p-value** smaller than **α**?

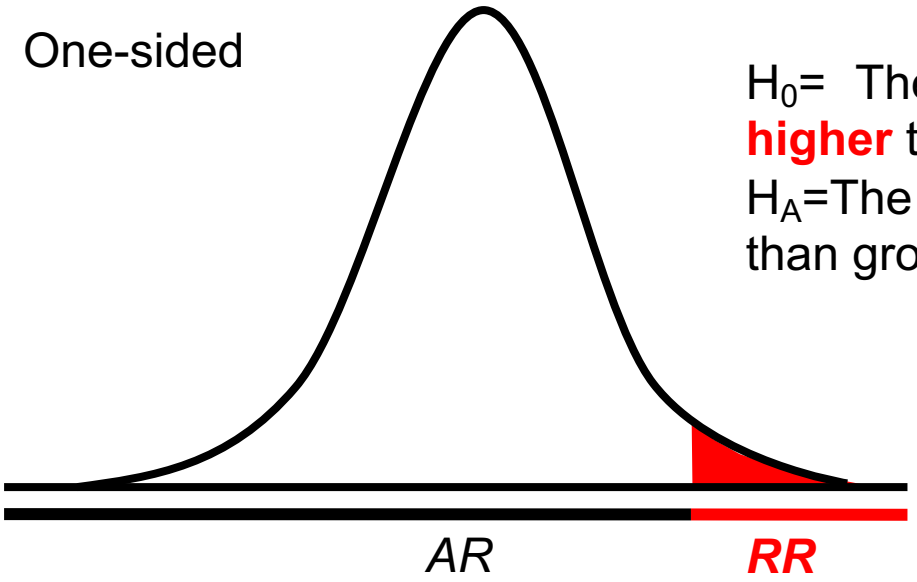






One and Two-sided hypothesis

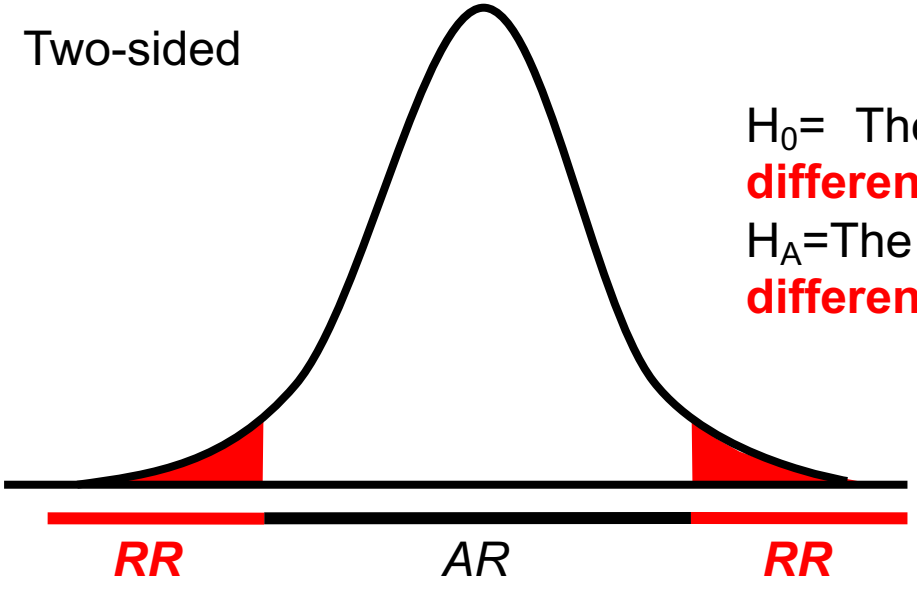
One-sided



H_0= The value of quantity in group A **is not higher** than group B

H_A=The value of quantity in group A **is higher** than group B

Two-sided



H_0= The value of quantity in group A **is not different** from group B

H_A=The value of quantity of interest in group A **is different** from group B

One and Two-sided hypothesis

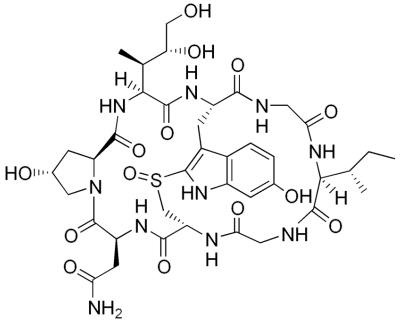
The **hypothesis** must be drawn **before starting the actual experiment**!!

e.g.
We want to know if **treating cells with α-amanitin** influences **gene expression**. (we don't know the effect of α-amanitin before starting the exp).

H_0 = **α-amanitin** doesn't affect gene expression in cells.

H_A = **α-amanitin** does affect gene expression in cells.

Amanita phalloides



Blocks Polymerases activity

The treatment would decrease the whole gene expression at once. In fact, the data would indicate us that there is a strong decrease. Yet, using a bias approach would lead us to **completely different test statistic and result**.
e.g. H_0 = α-amanitin doesn't increase gene expression in cells...

Comparing two groups

- Independent variable has two levels (Treatment -> Treated, Not Treated)
- Dependent variable (what we are actually interested in: RNA levels, enzymatic activity...)
- Biological and technical replicates (3 biological replicates and 3 technical)

RNA level	Treatment	Bio.Rep	Tech.Rep
10.6	No Treated	1	1
10.2	No Treated	1	2
9.4	No Treated	1	3
8.9	No Treated	2	1
9.3	No Treated	2	2
9.1	No Treated	2	3
10.3	No Treated	3	1
12.1	No Treated	3	2
9.7	No Treated	3	3
5.6	Treated	1	1
4.3	Treated	1	2
4.7	Treated	1	3
4.5	Treated	2	1
5.5	Treated	2	2
3.8	Treated	2	3
4.9	Treated	3	1
5.2	Treated	3	2
3.4	Treated	3	3

Comparing two groups

- Does the treatment **influence** gene expression?
- Is the difference "**significant**"?
- **H_0** = there is **NO difference** between the treatments
- **H_A** = there is **difference** between the treatments
- How likely is such a group means difference **occurring by chance**?

Which test should I use?

•depends on **the question** you are asking (hypothesis)

e.g.

B is higher than A, B is different from A...

•depends on the number of **independent** (cause) and **dependent** (effect) variables

•depends or

e.g.

Treatme

Genoty

•depends or

e.g.

Values

Years =

Behavior = Aggressive, Friendly

• SAMPLING HAS TO BE INDEPENDENT!

• SAMPLES HAVE TO BE REPRESENTATIVE OF THE POPULATION!

•depends on the **assumption of the test**

e.g. Is the distribution normal?

are variances equal?

...

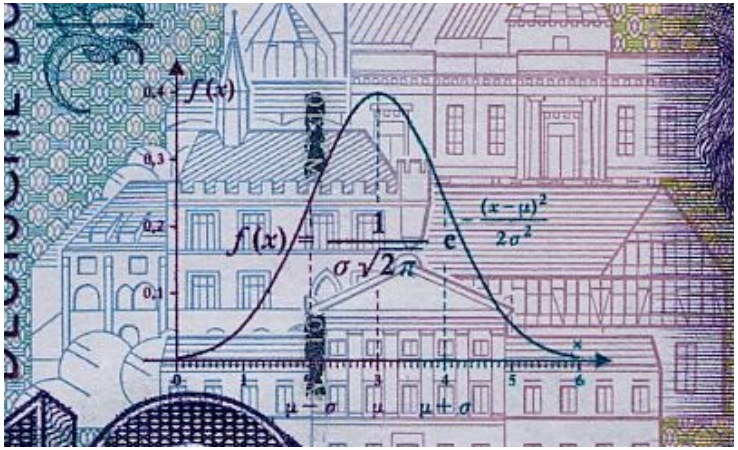
T-test (unpaired, two-tailed)

W.S. Gosset

REQUIRMENTS:

1. **Normally distributed** values
2. **Equal variances** in both groups
1. **Group size** can be **different**
2. **Sample** were obtained **independently**

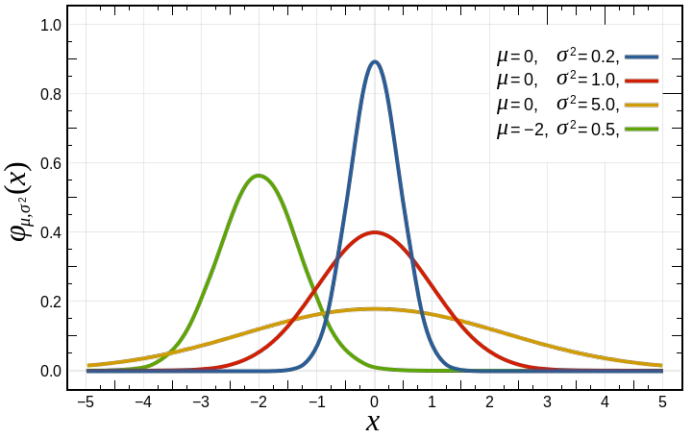
Normal distribution

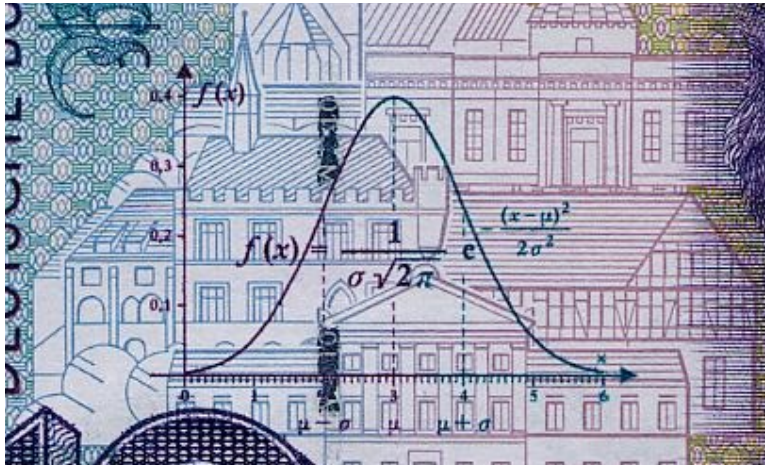


Normal distribution

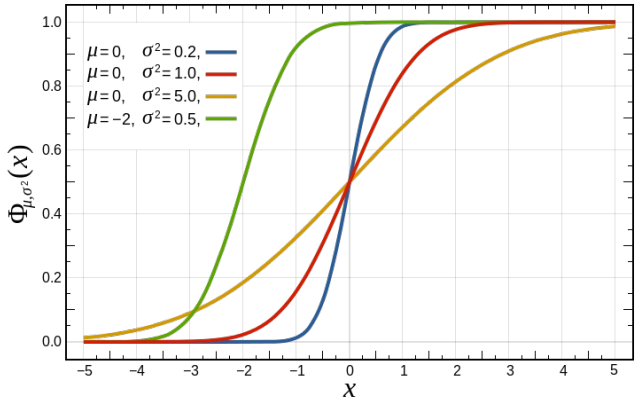
Normal distributed

it has a **probability density function** that follows the so called *“bell curve”*



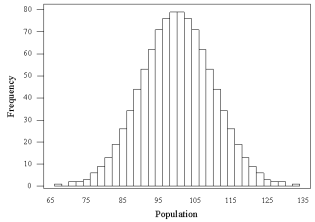


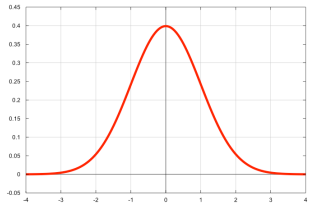
it has a specific **cumulative distribution**

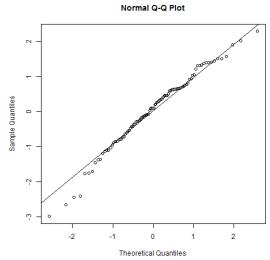


Normal distribution

How to test for normality??







Visual approaches:
Histogram, Density plot, Q-Q plot
It works when your N is high enough (N>>20)

Shapiro-Wilk test
H_0 = the distribution **is** normal
H_A = the distribution **is not** normal
Be careful on the p-val you get!!

```
> shapiro.test(heisenberg$HWWIchg)

        Shapiro-Wilk normality test

data:  heisenberg$HWWIchg
W = 0.9828, p-value = 0.0006497
```

***p-val* smaller than 0.05**

↓

we **reject H_0**

↓

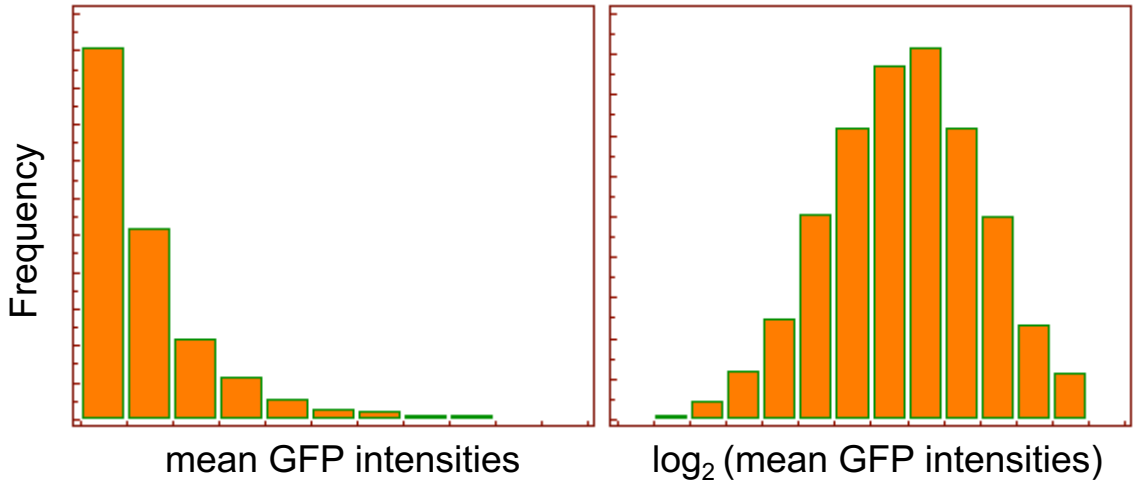
we **accept H_A**

Normal distribution

Certain population can be assumed to be **normally distributed** because of the ***central limit theory***

e.g.
we want to measure the size of *Macrochelys temminckii* beak.
*We can assume that the **population** will have a **normal distribution** when **N** is high enough. Therefore we can **treat our data as normally distributed**.*

Transform the data! *(Fluo intensities or qPCR data)*



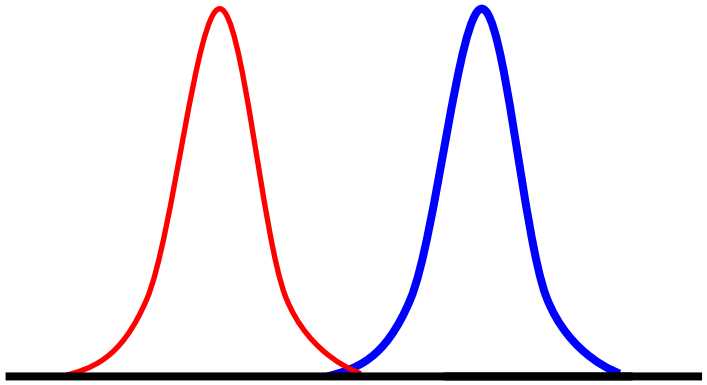
Other kind of transformation (rarely used)

Square-root transformation. This consists of taking the square root of each observation. The back transformation is to square the number. If you have negative numbers, you can't take the square root; you should add a constant to each number to make them all positive. The square-root transformation is commonly used when the **variable is a count of something**, such as bacterial colonies per petri dish, blood cells going through a capillary per minute, mutations per generation, etc.

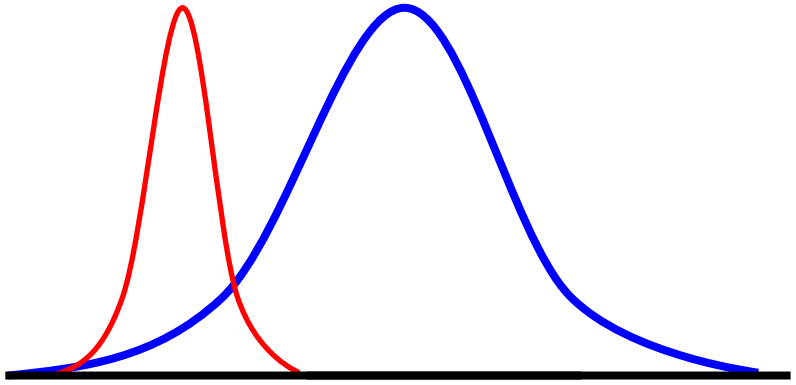
Arcsine transformation. This consists of taking the arcsine of the square root of a number. (The result is given in radians, not degrees, and can range from $-\pi/2$ to $\pi/2$.) The numbers to be arcsine transformed must be in the range -1 to 1. **This is commonly used for proportions**, which range from 0 to 1, such as the proportion of cells in culture that are infested by a mycoplasm.

Test on variances

Homoscedasticity:
Two distribution have the same scatter



Heteroscedasticity:
Two distribution have different scatter



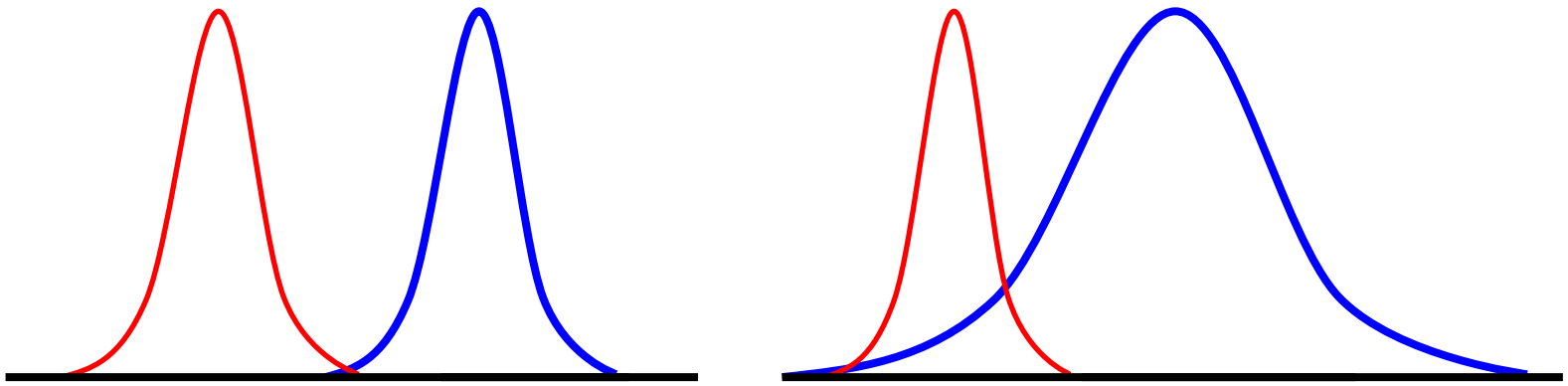
F-Test
H_0 = variances **are** equal
H_A = variances **are not** equal
PROs: High power on normal distributions
CONs: Too sensitive to the assumption of normality! You need high N!

Bartlett's test
H_0 = variances **are** equal
H_A = variances **are not** equal
PROs: High power on normal distributions
CONs: Applied on non-normal distributions will test for normality

Levene's test
H_0 = variances **are** equal
H_A = variances **are not** equal
PROs: High power on non-normal distributions
CONs: Applied on normal distribution is lesser powerful than Bartlett's test

t-test with Welch correction

I have a **normal distribution,** but I am not sure whether the **variances** are **equally distributed**

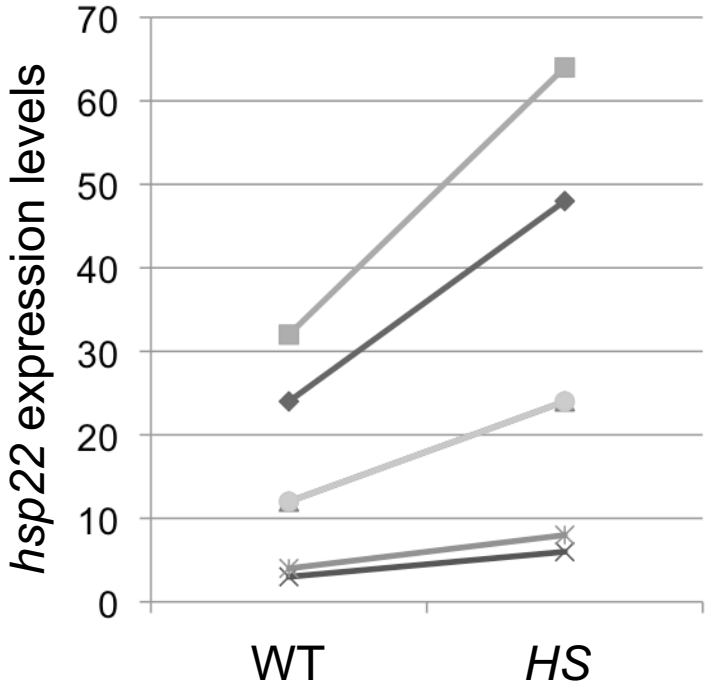


Welch's correction **allows for unequal variances** (Heteroscedasticity) and it is still reliable on distributions that do have equal variances (Homoscedasticity)

Another example to get somewhere else...

We want to know whether the mitochondrial **gene *hsp22*** in *D. melanogaster* is **activated during heat-shock response**.

	***hsp22* expression**	
Sample	**WT**	***HS***
S1	24	48
S2	32	64
S3	12	24
S4	3	6
S5	4	8
S6	12	24



Two-tailed t-test with **Welch** correction

H_0 = There is **no difference**
H_A = There is **difference**

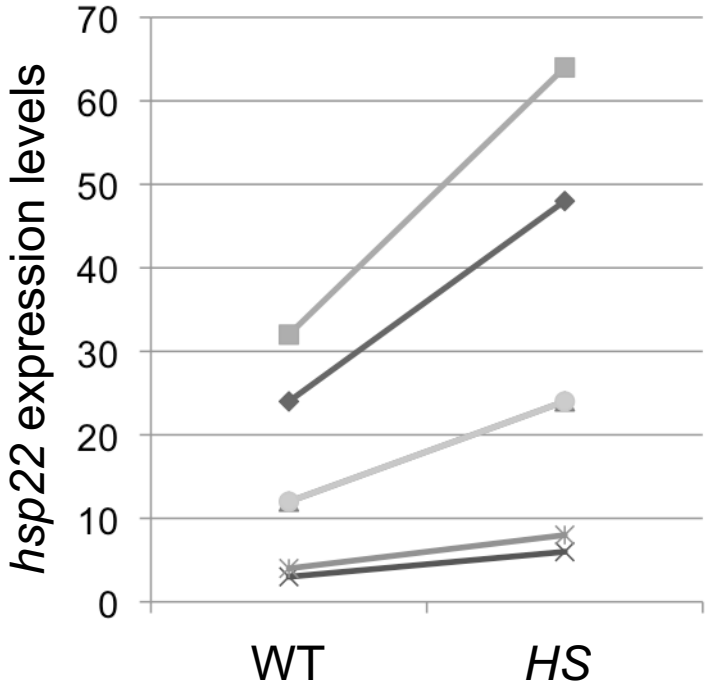
p-val = 0.194, we **accept H_0**

PAPER TITLE: ***The mitochondrial gene hsp22 is not associated with the heat-shock response in D. melanogaster***

Another example to get somewhere else...

We want to know whether the mitochondrial **gene *hsp22*** in *D. melanogaster* is **activated during heat-shock response**.

	***hsp22* expression**	
Sample	**WT**	***HS***
S1	24	48
S2	32	64
S3	12	24
S4	3	6
S5	4	8
S6	12	24



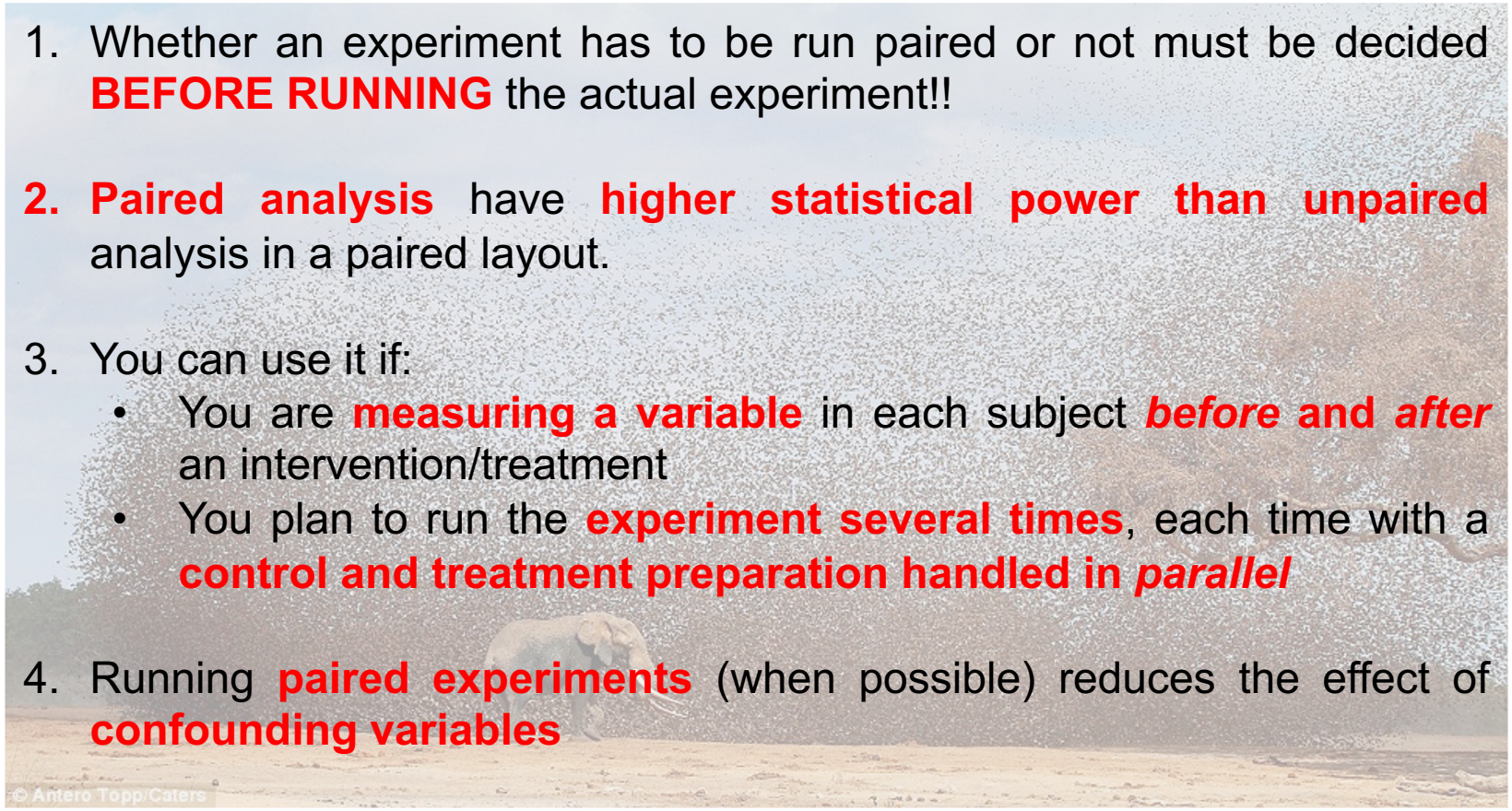
Two-tailed paired t-test with **Welch** correction

H_0 = There is **no difference**
H_A = There is **difference**

p-val = 0.027, we **reject H_0 and accept H_A**

PAPER TITLE: ***The mitochondrial gene hsp22 is activated upon heat-shock induction in D. melanogaster***

The importance to be paired



1. Whether an experiment has to be run paired or not must be decided **BEFORE RUNNING** the actual experiment!!

2. **Paired analysis** have **higher statistical power than unpaired** analysis in a paired layout.

3. You can use it if:
   - You are **measuring a variable** in each subject ***before* and *after*** an intervention/treatment
   - You plan to run the **experiment several times**, each time with a **control and treatment preparation handled in *parallel***

4. Running **paired experiments** (when possible) reduces the effect of **confounding variables**

Power analysis of two-tailed unpaired t-test

1. **Sample size**
2. **Effect size** (the strength of a phenomenon) $\theta = \frac{\mu_1 - \mu_2}{\sigma},$
   - pioneer experiments will give us this information
3. **Significance level α** (usually set at 0.05)
4. **Power**, 1-β (the probability of **False Positive – Type I Error**) – typically set to 80% or 90%

```
Two-sample t test power calculation
              n = 4.574784
              d = 2.5
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
NOTE: n is number in *each* group
```

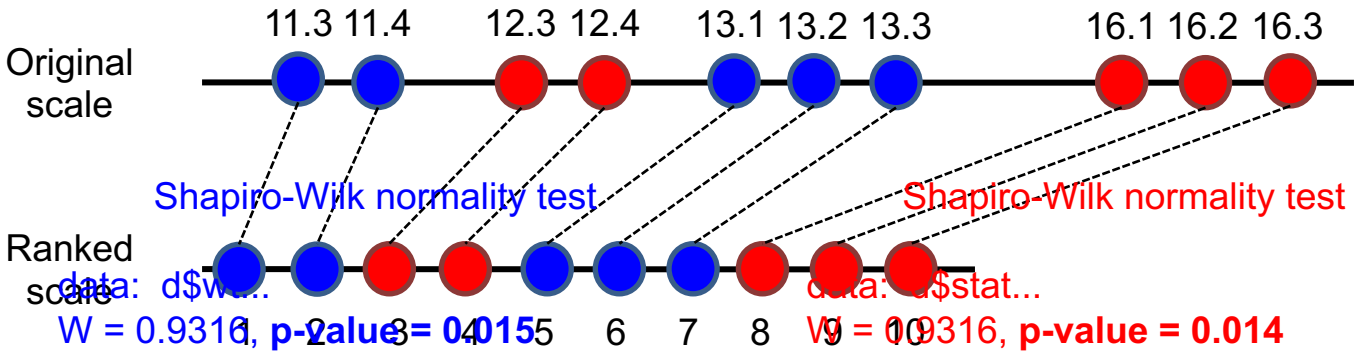
Heinrich Heine Universität Düsseldorf – Online resource:

http://www.gpower.hhu.de/

Non-normal distributions

We want to know if ***STAT5a*** influences granulocytes' size

	Cell diameter [µm]				
WT	13.3	11.4	11.3	13.2	13.1
STAT5a RNAi	16.3	12.4	12.3	16.2	16.1



Rank Sum **Group 1 (WT)**:
1+2+5+6+7= 21

Rank Sum Group 1 (*STAT5a*):
3+4+8+9+10 = 34

Wilcoxon rank sum test

data: d$WT and d$stat...
W = 1, p-value = **0.22**
alternative hypothesis: true location shift is not equal to 0

→ **t-test**
p-value = 0.08

Wilcoxon-test on normal distribution

For large data set (N > 50) a wrong decision does **NOT matter**.

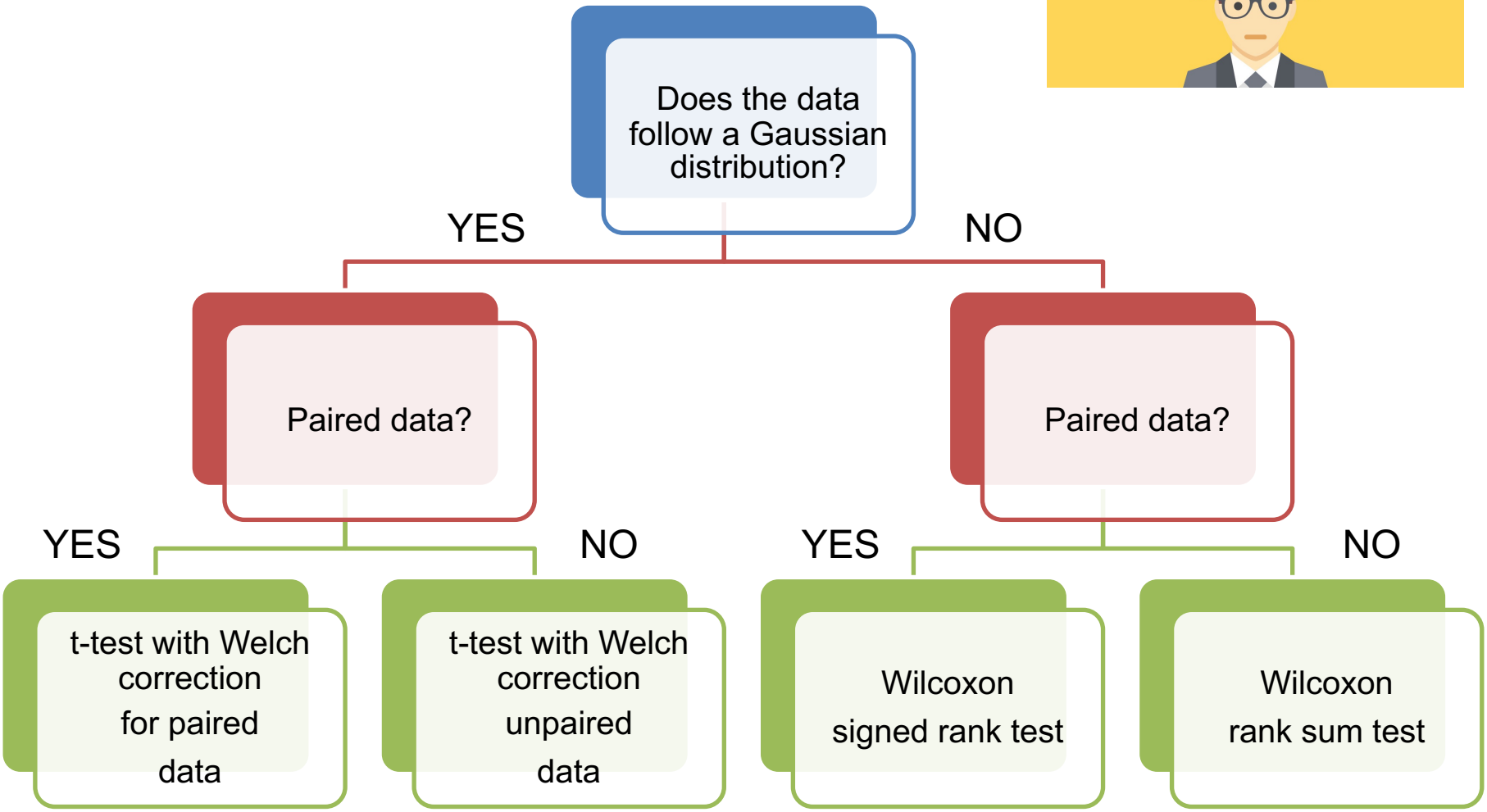
Wilcoxon-test has the **95% of the t-test power** when applied **to normal distributions**

For **small data set** the wrong choice **matter**

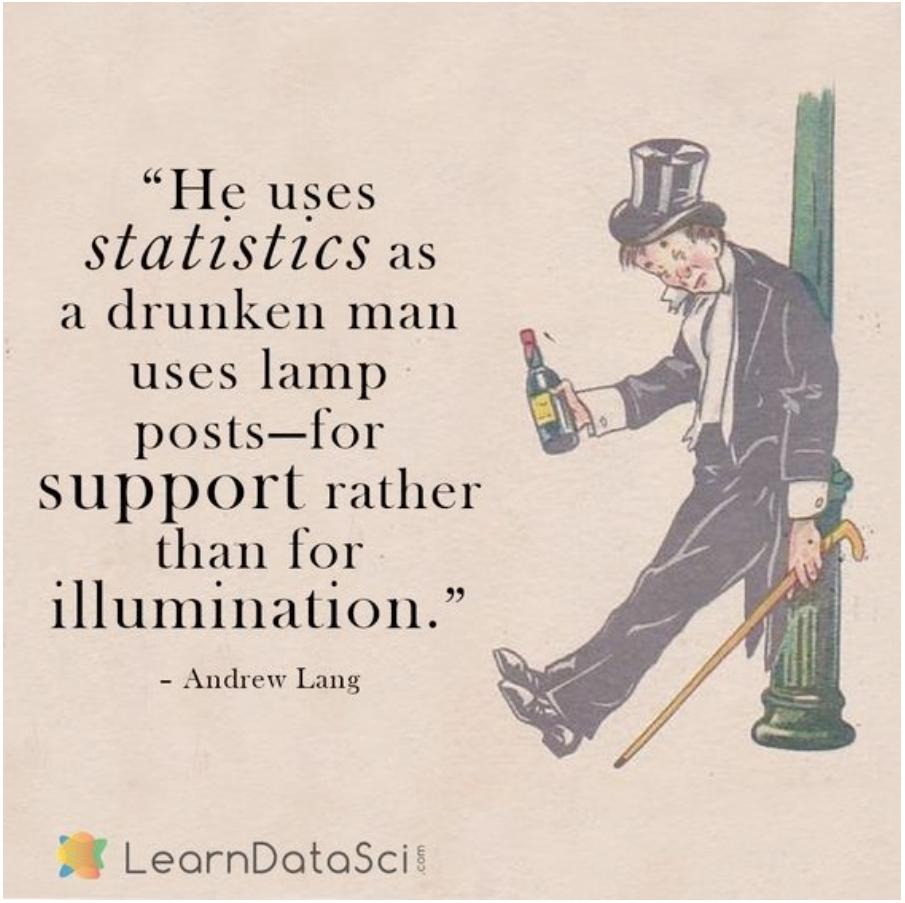
non-parametric test (e.g. Wilcoxon-test) have a **low power**

parametric test are **not robust**

Comparing 2 groups of continuous data

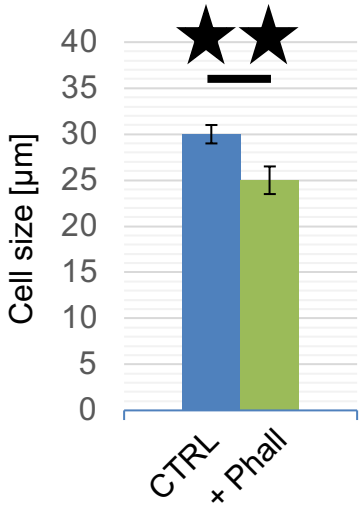


STATISTICAL SIGNIFICANCE doesn't necessarily mean **RELEVANCE**



What the *p-val* is **NOT**

We want to investigate whether Phalloidin influences cell growth.

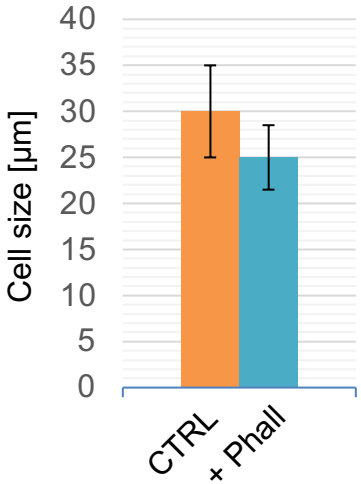


n = 30

Amanita phalloides

Two-sided t-Test with Welch's correction

p-val = 0.003



n = 3

Two-sided t-Test with Welch's correction

p-val = 0.19

What the *p-val* is **NOT**

"There is the 0.3% probability that the two means are equal, and 99.7% that are different."

No true! The p-val indicates that **if the H_0 is true** we would see a **difference between the two means 0.3% of the time.** The *p-val* is indicative of the odds of seeing the difference, is not telling you anything *directly* on what you have measured.

"The low *p-val* indicates that the alternative hypothesis is true."

The **low *p-val*** is only telling us that **we reject the H_0** not that the H_A is true. We accept the **H_A** with a **certain probability of committing an error (type I error).**

"The high *p-val* proves that there is not effect."

Nope…**absence of evidence is not evidence of absence** *(Altman DG, Bland JM 1995).* The high *p-val* indicates that we cannot reject the H_0. The missed difference is only a matter of the sample size. (*Ranstam J. 2012)*

"Such low *p-val* indicates an important difference between the two means."

No true! The **low *p-val*** tells you only that the **two means are significantly different**. The ***p-val* proves that such difference does exist**, but it says **nothing about the actual/practical relevance** of your data.

Real troubles with p-values (who cares about stars)

1. **p-values** do not reveal the underlying **effect size**
2. **p-values** scale with **n**
3. for **large N small effects may become significant**
4. for **small N** the observed effects might be relevant but **not statistically significant**
5. **p values** have **strong variation** between repeated experiments
6. **p-value is not** necessarily a proxy for **reproducibility**
7. **Statistical significance does not** tell anything about **biological relevance**

How to report p-values in your papers.

1. Report the test applied and the parameters you used
    - e.g. two-tailed t-test with Welch's correction on paired data
2. Try to avoid terms in the text such as "*extremely significant*" or "*statistically significant*"...use only **significant.**
3. It is preferred to report the **actual value of the p-values** rather than categorize them (such as $p < 0.05$ etc).
4. Always report **N** (biological) and then the ***p-values***

"Statistically significance is neither a necessary nor a sufficient condition for providing a scientific result"- S. Ziliak

Using the p-val for *within-experiment analysis* does NOT tell you anything about the biological robustness of your sample.

e.g.

ChIP-seq peak calling

Database searches

Peptide identification in mass spectrometry

Reporting *p-val* on *technical replicates* is not an inferential information. We cannot conclude much about the actual *biological population*.

Multiple testing – Bonferroni correction

Frank wants to know whether the EMF complex in plants changes components during veranlization. To do so, Frank pulls down CLF, the core component of such complex, and with it all the proteins that are associated.

	log_2 Fold change	***p-val***
CLF	0.2	0.8
EMF2	-2.3	0.002
FIE	-0.45	0.03
MSI1	0.23	0.06
VRN2	1.12	0.001

Bonferroni adjustment: α/k

$\alpha = 0.05 \rightarrow \alpha_{adj} = 0.01$

A p-val of 0.05 means that there is a 5% chance of getting your observed results, if the H_0 were true.
If we repeat the test n times, we change this probability

$\alpha_k = 1-(1-\alpha)^k$

Although we keep the α at 0.05 per each comparison (k), we drastically increase the probability of false positive. In this example is 22%.

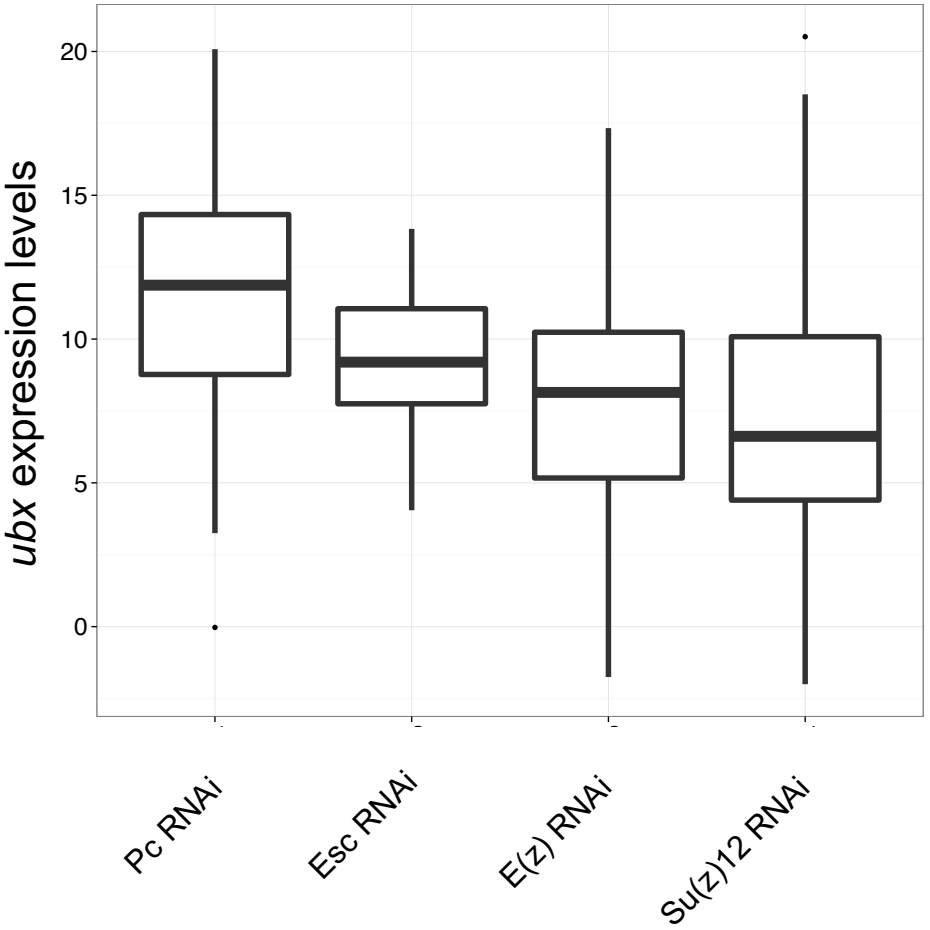
ANOVA

H_0 = means of the measurement variable are **ALL** the same for the different categories of data

H_A = the means of the measurement variable are **NOT ALL** the same

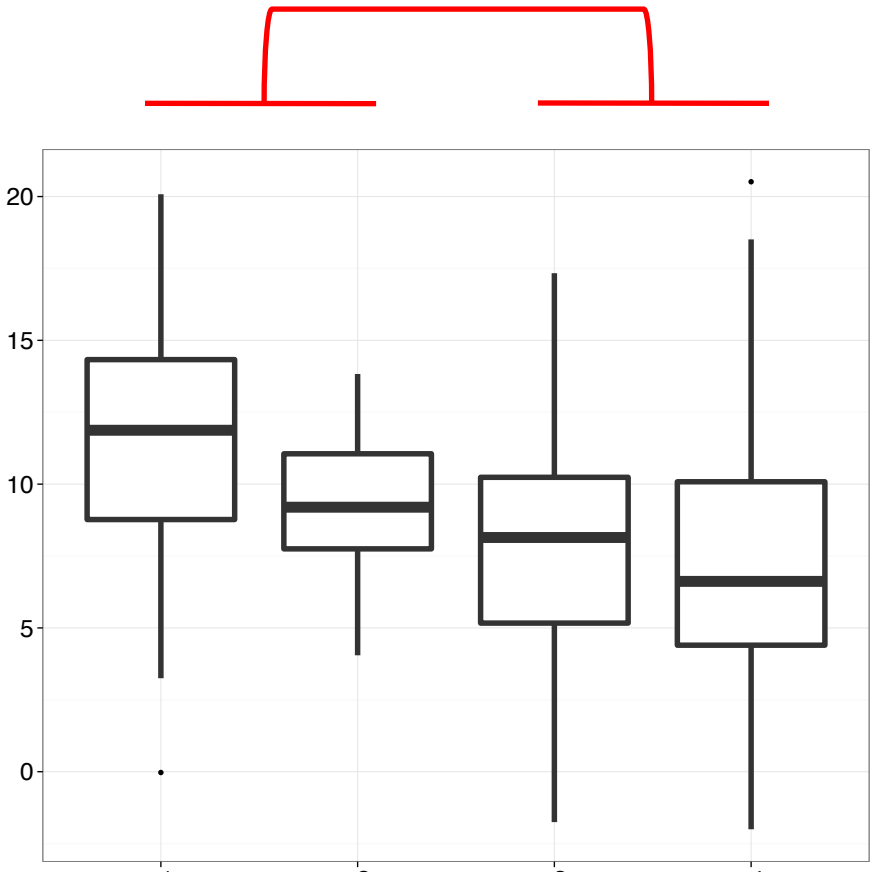
e.g.
We want to know whether *ubx* expression is equally affected by the depletion of its major repressors.

```
> summary(aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
df$name      3  403.1  134.37   8.739 1.81e-05 ***
Residuals  196 3013.6   15.38
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



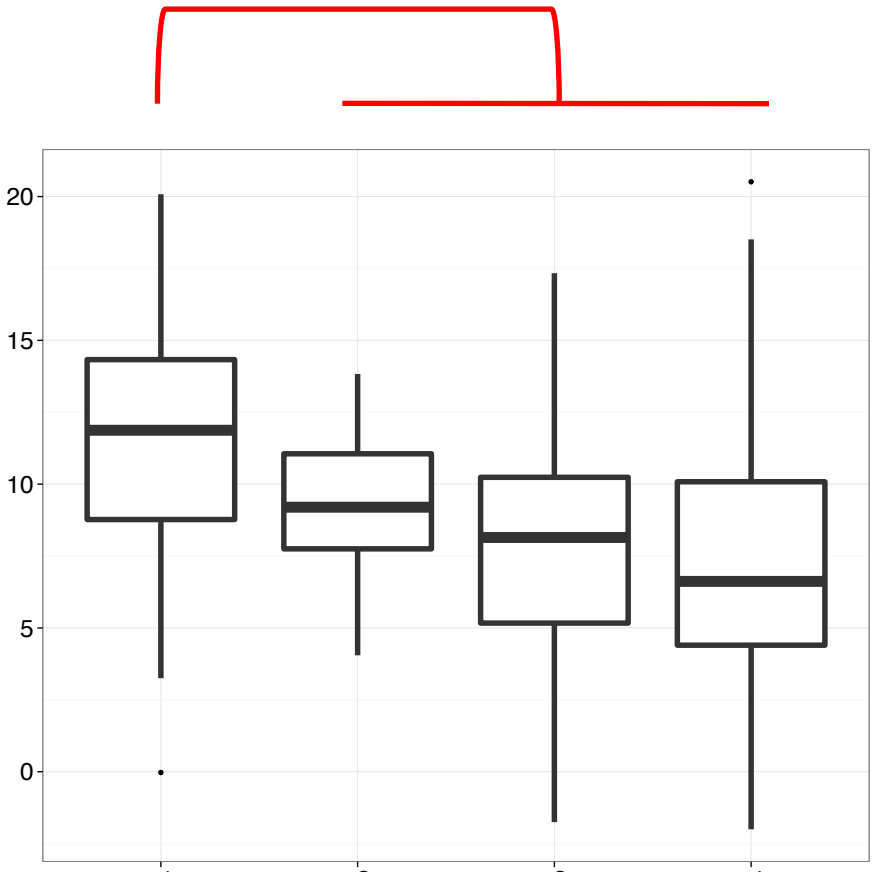
ANOVA – is it enough? Post Test

It depends on the question you have...



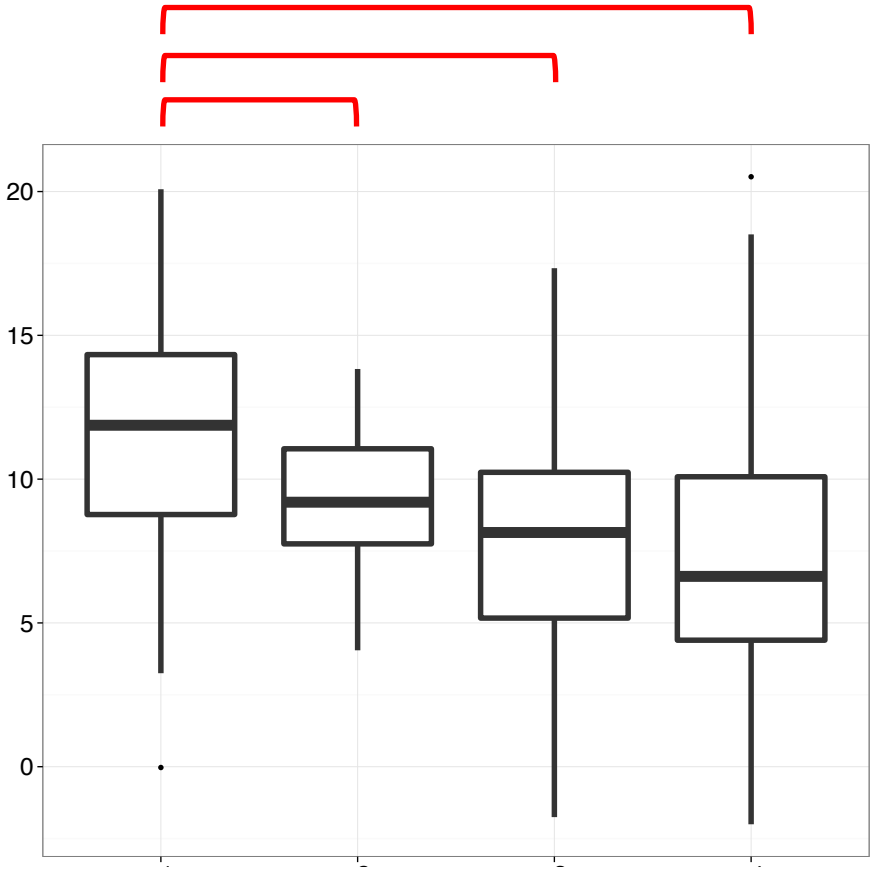
ANOVA – is it enough? Post Test

It depends on the question you have...



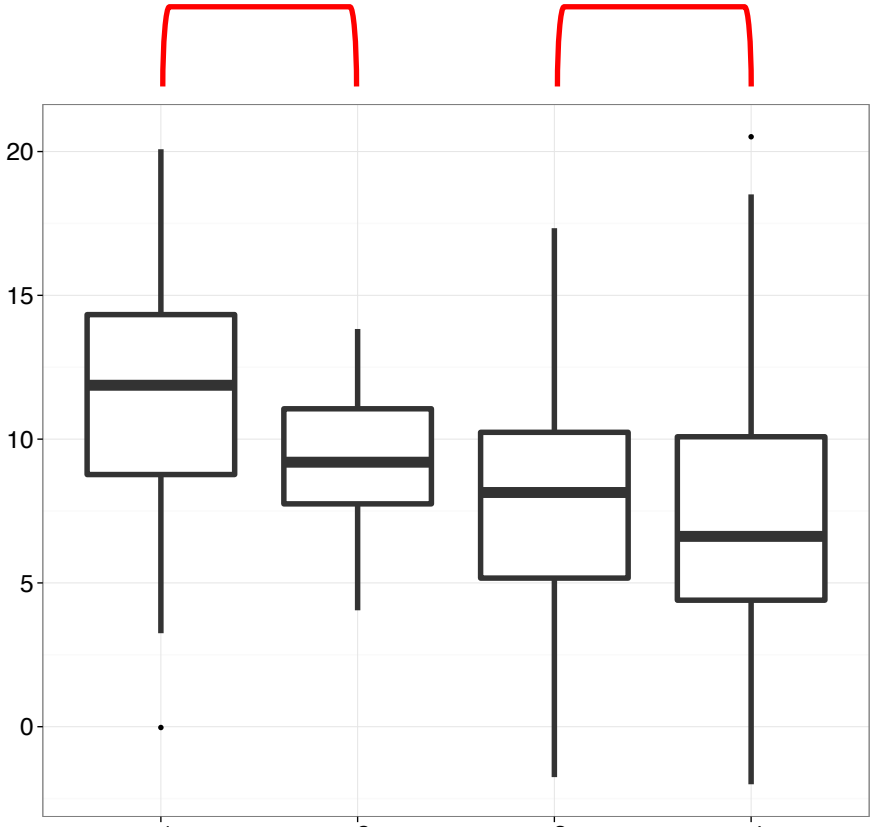
ANOVA – is it enough? Post Test

It depends on the question you have...



ANOVA – is it enough? Post Test

It depends on the question you have...



Planned Post Test

Orthogonal comparison – all the comparison are independent.

no p-val adjustment needed

just pool the data for another I-way ANOVA leaving the groups outside the comparison.

Non orthogonal comparison – the comparison depends to each other.

specific method must be applied: ***Dunn-Sidák*** or ***Bonferroni***

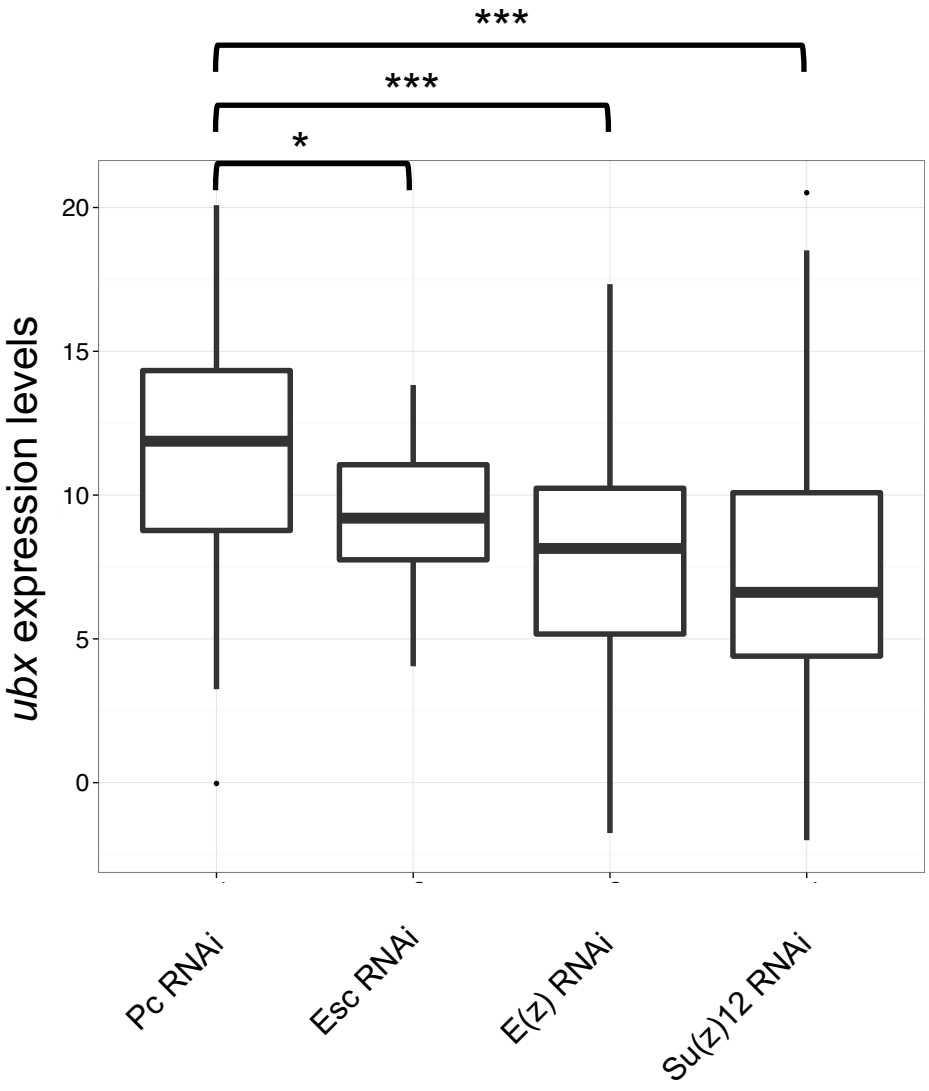
Tukey – all the pairs

> TukeyHSD(aov)
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = df$val ~ df$name)

$`df$name`

	diff	lwr	upr	p adj
Esc-Pc	-2.0542388	-4.086349	-0.02212892	**0.0464661**
E(z)-Pc	-3.3221313	-5.354241	-1.29002141	**0.0002036**
Su-Pc	-3.5960749	-5.628185	-1.56396500	**0.0000475**
E(z)-Esc	-1.2678925	-3.300002	0.76421742	0.3715856
Su-Esc	-1.5418361	-3.573946	0.49027383	0.2044745
Su-E(z)	-0.2739436	-2.306054	1.75816632	0.9853263



Tukey – all the pairs

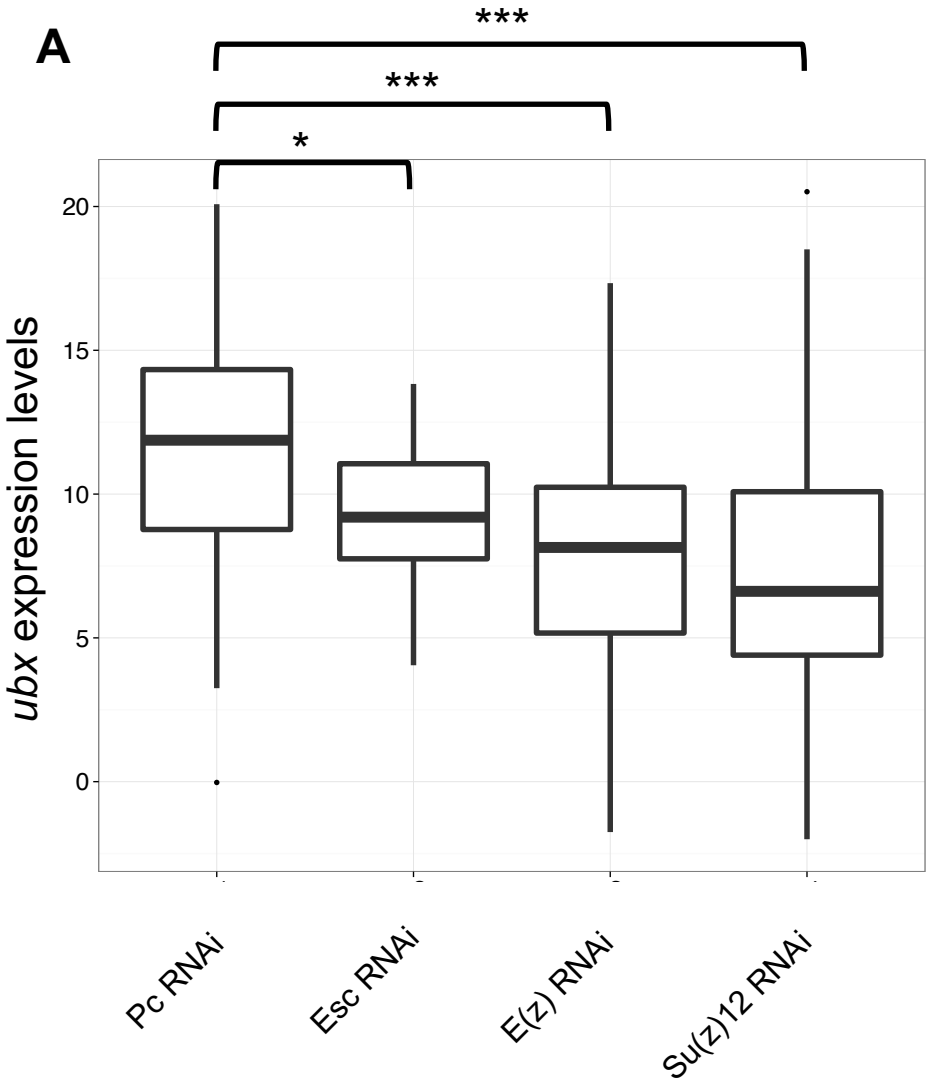


Figure 1A. Levels of *ubx* expression upon induction of specific RNAi constructs. One-way ANOVA (F=8.739, P=1.81e-05) followed by Tukey's-test.
(***p-val***: *Pc-Esc=0.0464661, Pc-E(z)=2.0e-04, Pc-Su(z)12=4.8e-5)*, n = 50. Non significance is not reported

Planned Post Test

1. **Compare all pairs**: **Bonferroni**, **Tukey**, Student-Newman-Keuls, preferred method depends on number of groups

2. **Compares a set of treatments against a single control mean: Dunnett**

3. **All possibilities** (contrasts): **Scheffé** test (low power)

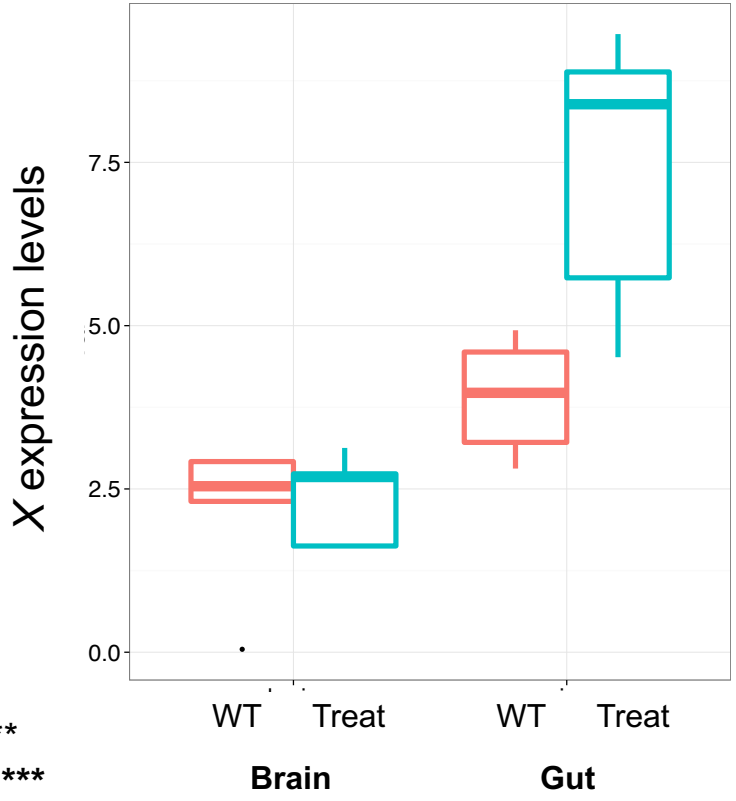
4. **Groups naturally ordered**: **test for trends**

Two-way ANOVA

1- **First factor** differences of means

2- **Second factor** differences of means

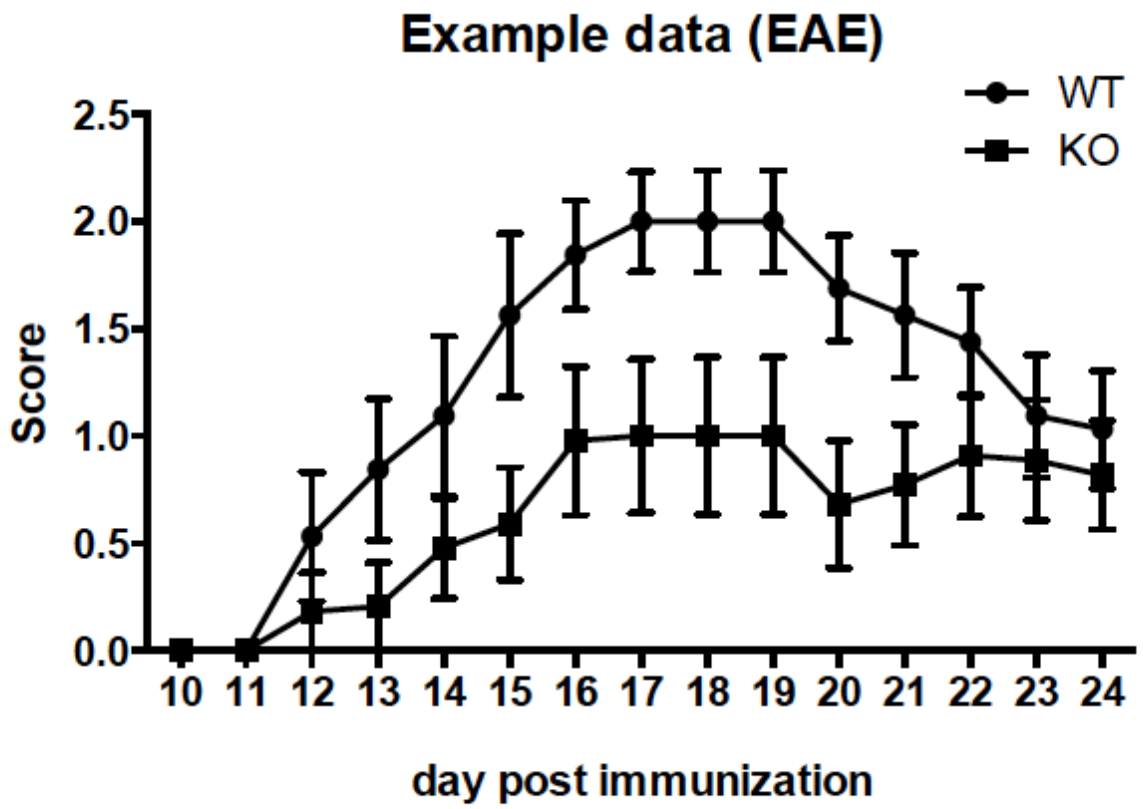
3-Interaction between **Fact 1** and **Fact 2**



> summary(aov)

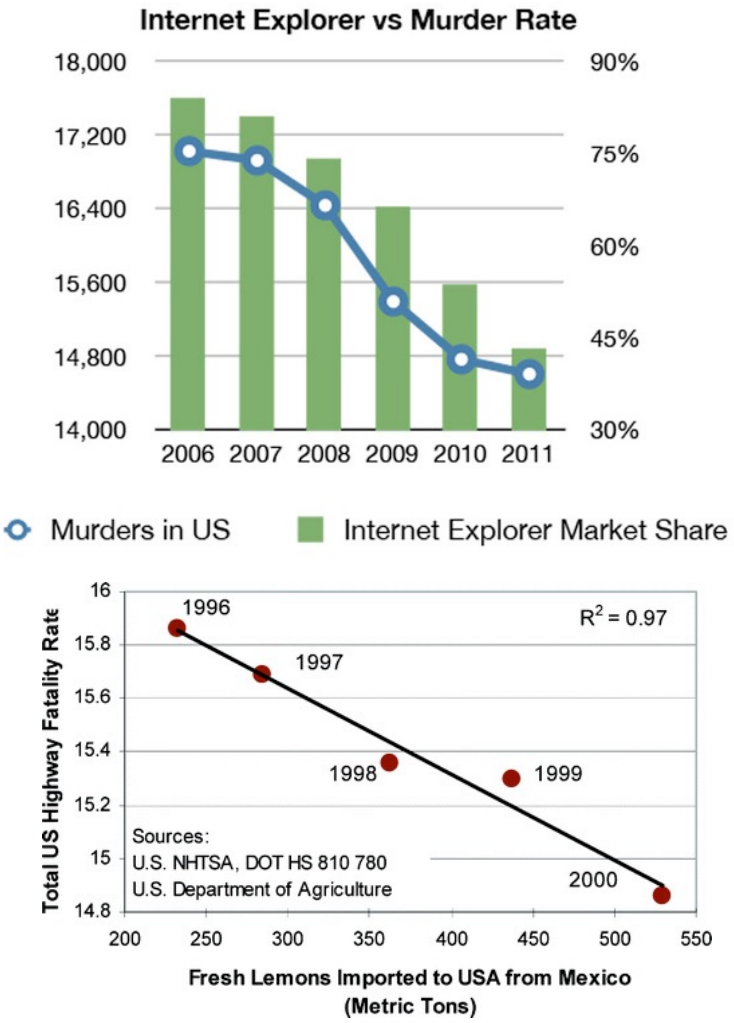
	Df	Sum Sq	Mean Sq	F value	**Pr(>F)**
df$name	1	17.09	17.09	9.268	**0.00773** **
df$tiss	1	57.67	57.67	31.276	**4.05e-05** ***
df$name:df$tiss	1	13.52	13.52	7.332	**0.01552** *
Residuals	16	29.50	1.84		

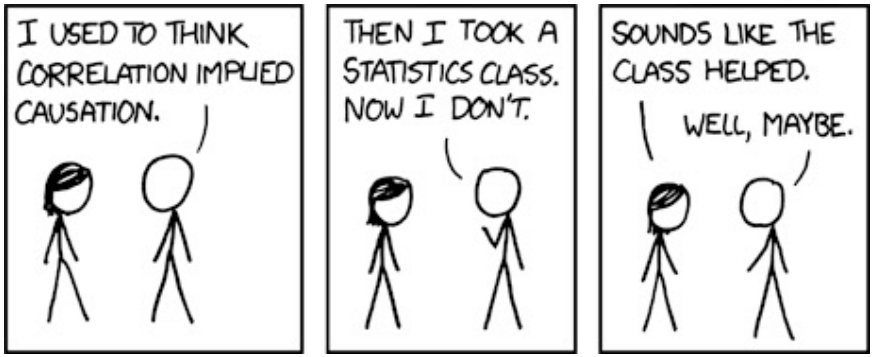
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

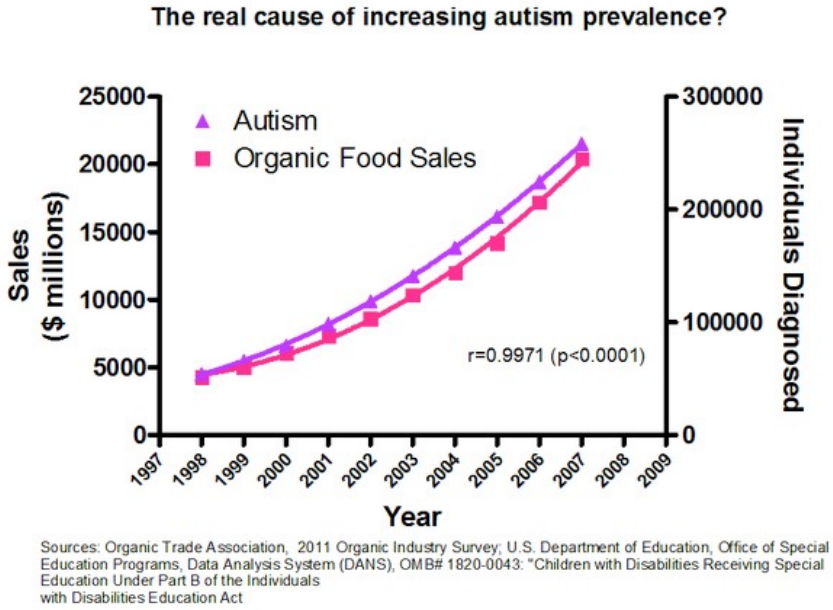

Example data (EAE)
WT
KO
Score
2.5
2.0
1.5
1.0
0.5
0.0
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
day post immunization

Bivariate Analysis

Correlation vs causation

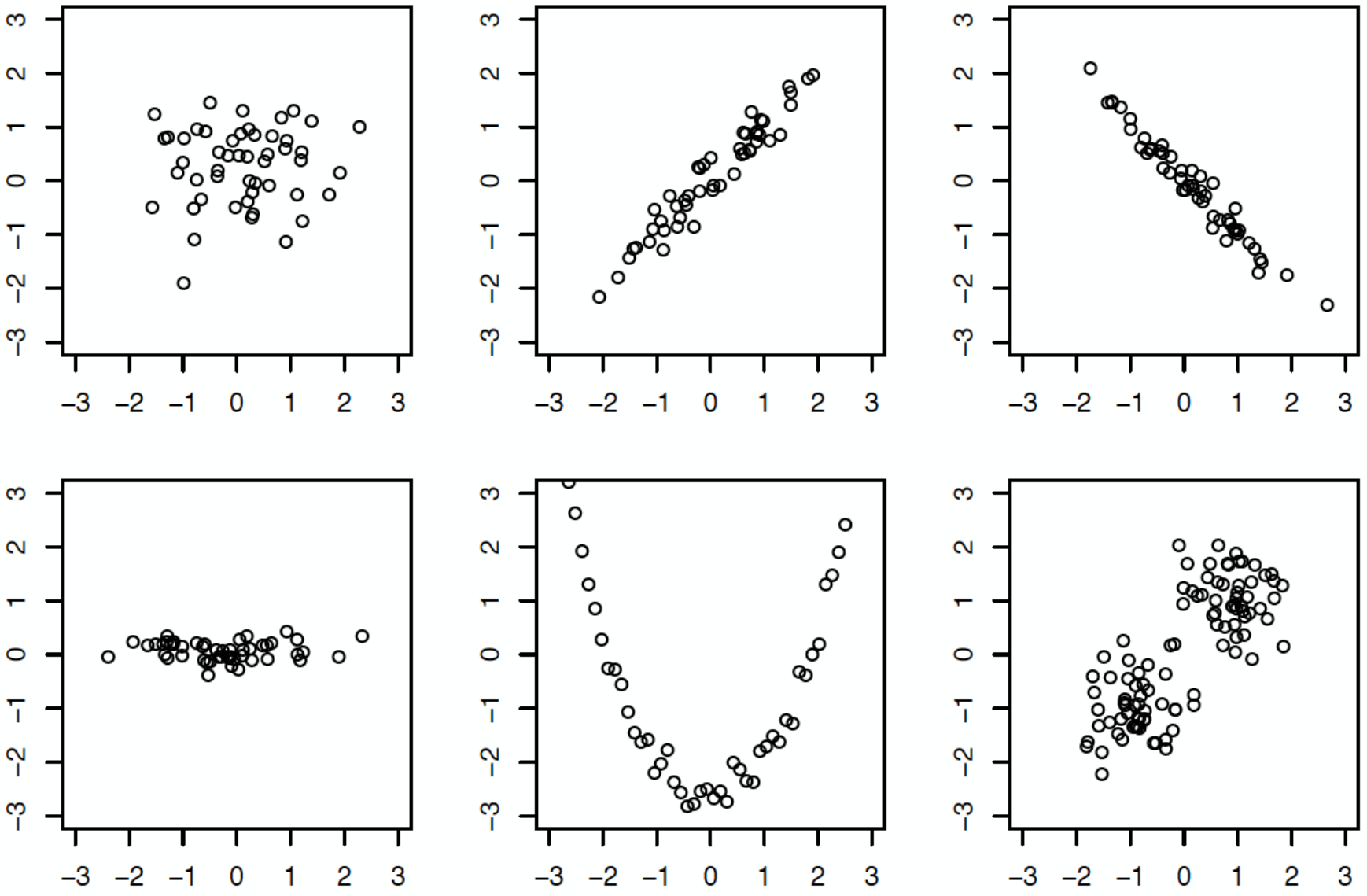






Correlation doesn't imply causation. After correlation has been measured, **causation must be proven** and not simply assumed.

Correlation between two variables



Pearson's correlation

1. Useful for **normally distributed variables** (but not only for those)
2. Measures the **degree of linear dependence**
3. $-1 \geq r_{xy} \leq 1$
4. $r_{xy} = 1/-1$: perfect linear **dependence**
5. $r_{xy} = 0$: linear **independence**

The **significance of a correlation** is expressed in probability levels (p-*val*) telling how likely a given correlation coefficient will occur given no relationship in the population

Pearson's correlation vs Spearman's correlation

Pearson correlation is a mesure for **linear** dependence

Spearman correlation is a measure for **monotone** dependence

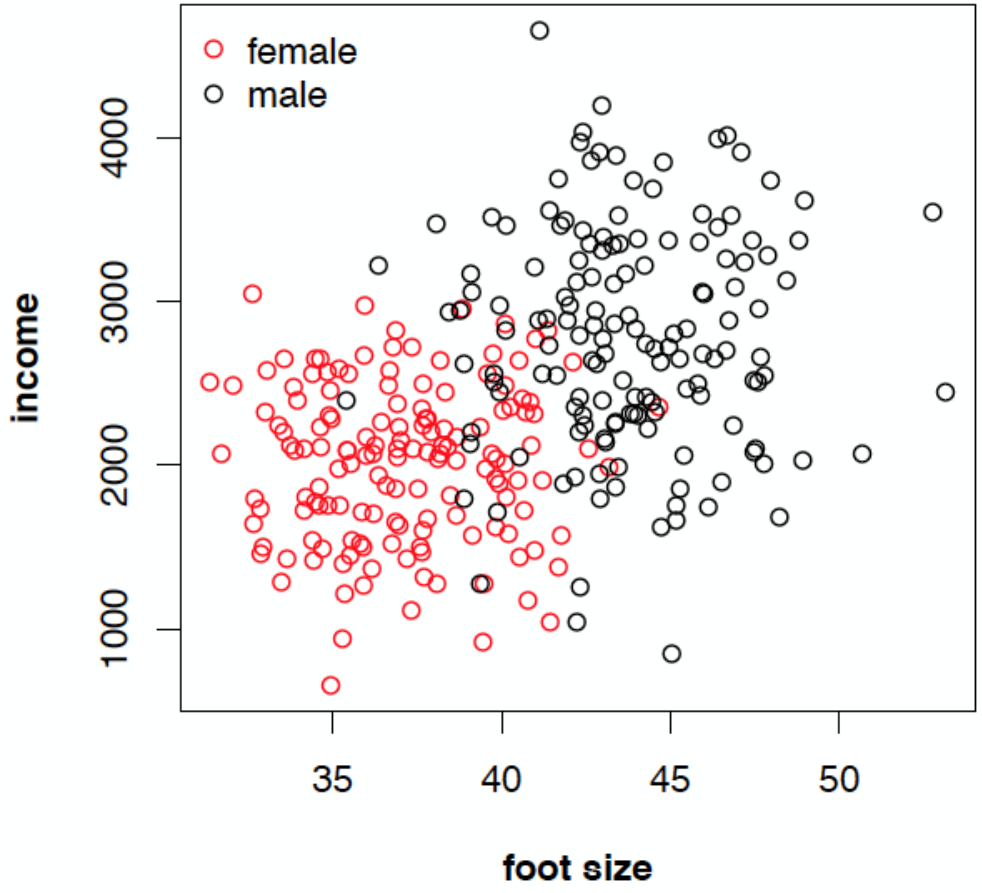
Spearman correlation is **less sensitive** than the Pearson correlation to strong outliers that are in the tails of both samples.

Correlation coefficients do **NOT** tell you anything about the existence (or non-) of **functional** or **causal dependence**.

The **significance of a correlation** is expressed in probability levels (p-*val*) telling how likely a given correlation coefficient will occur given no relationship in the population

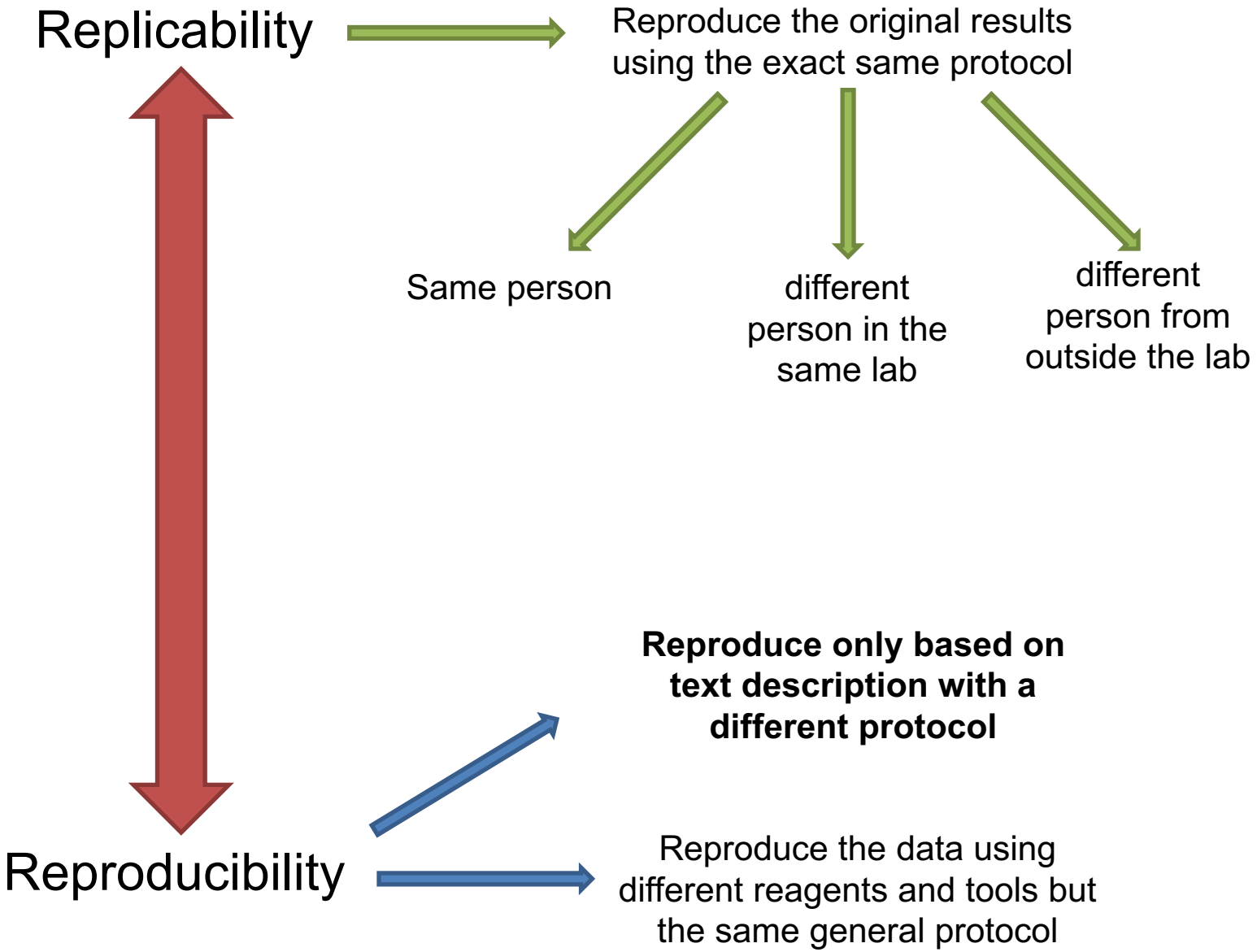
Confounding variables

A **confounding variable** is a variable, other than the independent variable you are interested in, that **may affect the dependent variable**.



Pearson correlation
r = 0.42

Responsible Research


Replicability
Reproduce the original results using the exact same protocol
Same person
different person in the same lab
different person from outside the lab
Reproducibility
Reproduce only based on text description with a different protocol
Reproduce the data using different reagents and tools but the same general protocol

Sampling bias – how to avoid it

1. **Blinding**

   You should handle *control* and *treated* in the same exact way. *In theory the person conducting the experiment shouldn't even known whether she/he is handling the control or the treated sample (blind).*

2. **Randomization**

   the sample should be assigned randomly to experimental groups

3. **Exclusion**

   It is allowed only when the criteria are set before the experiment starts.

4. **Confounding**

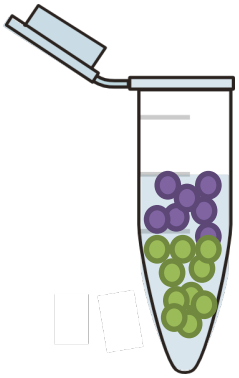
   Possible confounding factors should be identified before the data are actually acquired.

Blinding and randomization

Is it really possible?

The short answer **is not**, but you can get **close enough**.

The ultimate **aim** is to **not be bias** towards **a specific type of observation**



Pools of different type of cells can have different sedimentation properties (e.g. adipocytes tend to float). **Mixing a tube** before sampling the cells, it is a crucial step that **introduces randomization** in our data.

Ask a colleague:

e.g. **change label** to your data (making sure that you can track them back) in a way that the **new label** method **is not bias** by the sampling method you used. **Ask a colleague** to analyze your data if he can.

On-line Resources:

https://www.randomizer.org/

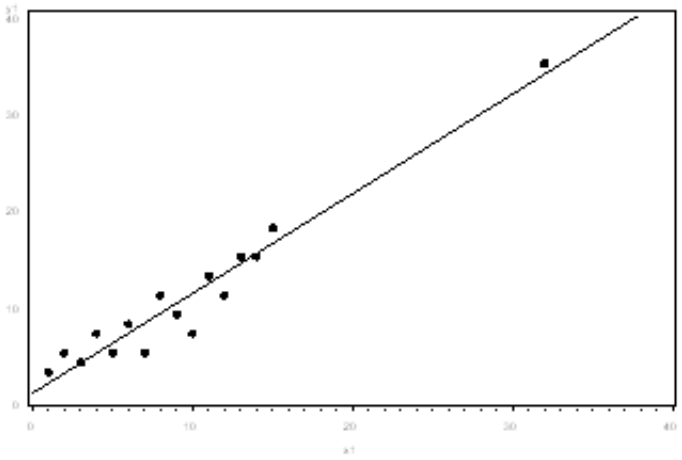
https://www.random.org/lists/

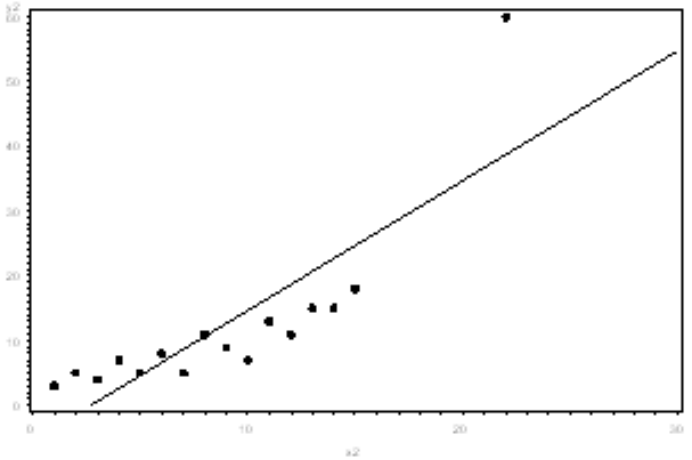
Exclusion

Exclusion parameters has to be decided **before** starting the actual experiment. It is very likely that you would be bias in front of your data once you have collected them. Making decision at this point is **very risky**.

e.g.
Data that will follow outside the whiskers in a boxplot will be excluded.

e.g.
Analysis of the variances' distributions from different replicates in order to remove "noisy samples" (technical issues).





Confounding - Examples

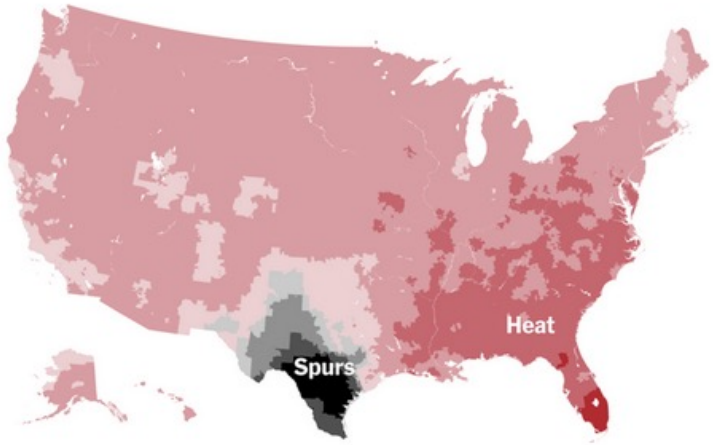
Different species of fireflies can produce different colored lights, which is in a range from green to yellow.
If we would measure the **glowing light-wave frequency** with a machine that **filters-out yellow**, we would conclude that fireflies glow only in green.

In higher eukaryotes, RNA nucleases usually **digest the RNA from the 3' to the 5'** (e.g. in *S. cerevisiae* is the other way round).
If we want to check **the differences in gene expression** by PCR, we should be **consistent with the region of the transcript we are amplifying**. Starting from the assumption that any pair of primers amplifies the same target equally, it could lead us to wrong conclusions.

How bad it can turn to be

vs

using **Facebook** data

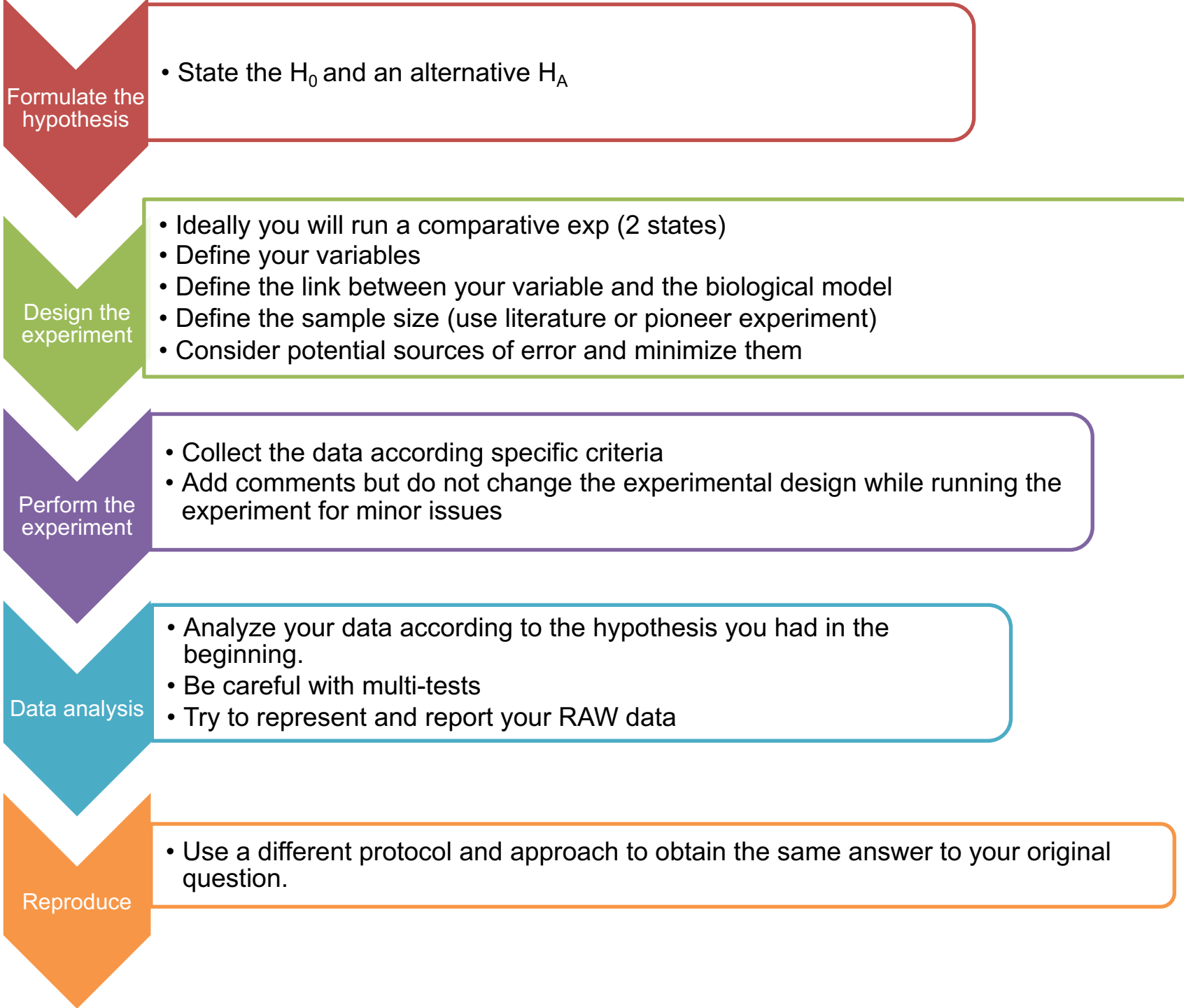


The Upshot – New York Times

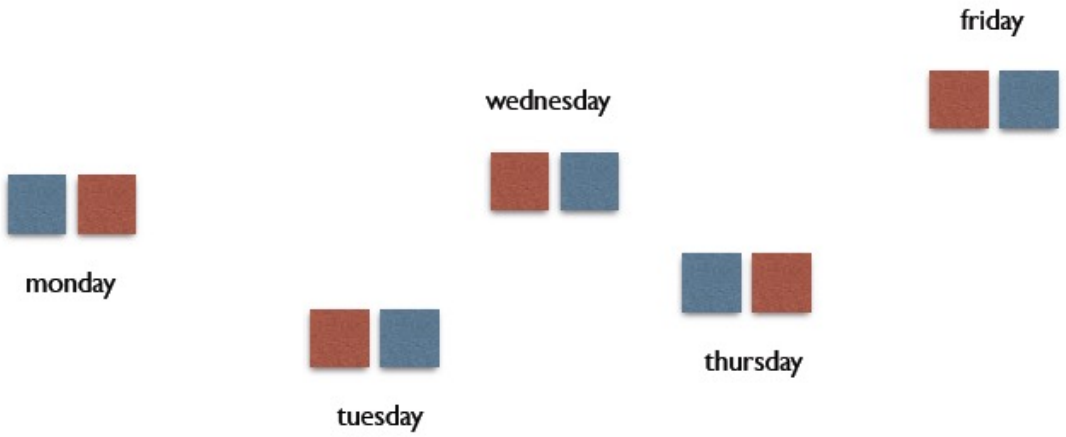
ESPN viewers

ESPN channel

The Ideal Experiment


Formulate the hypothesis
• State the H_0 and an alternative H_A
Design the experiment
• Ideally you will run a comparative exp (2 states)
• Define your variables
• Define the link between your variable and the biological model
• Define the sample size (use literature or pioneer experiment)
• Consider potential sources of error and minimize them
Perform the experiment
• Collect the data according specific criteria
• Add comments but do not change the experimental design while running the experiment for minor issues
Data analysis
• Analyze your data according to the hypothesis you had in the beginning.
• Be careful with multi-tests
• Try to represent and report your RAW data
Reproduce
• Use a different protocol and approach to obtain the same answer to your original question.

the ideal design



Randomized block design, only 2 factor levels (control *blue,* treatment *red*)

Suited to control for **day-to-day fluctuations** which are very common. **Change reagents, batches of cells...between blocks!**

Here the hypothesis could be tested by **paired t-test**...

N is too small, help!

Improve experimental design

- **simple comparative studies** (2 states) have higher power than complex studies
- **random blocking** will reduce systematic errors

Improve the power of statistical test

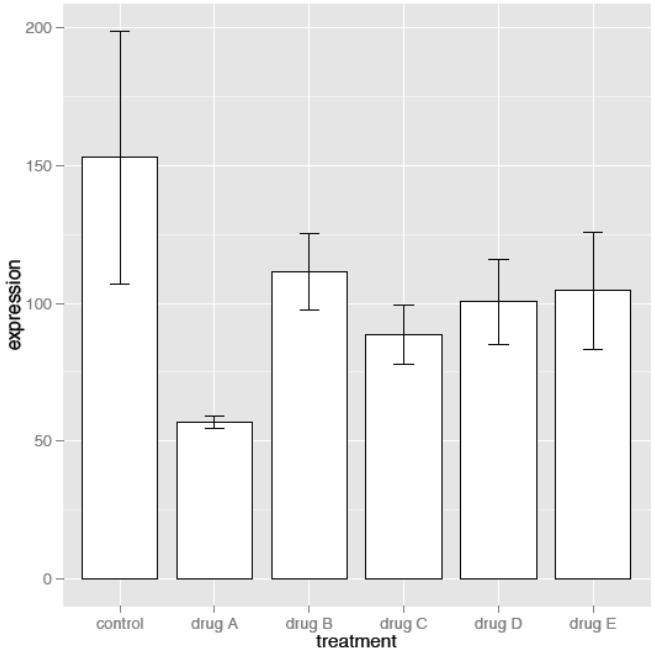
- **paired test** instead of unpaired
- avoid **making comparison** that are of **no interest**

Example

- test a couple of drugs on whether they affect the expression of a gene
- quick shot: qPCR with technical replicates

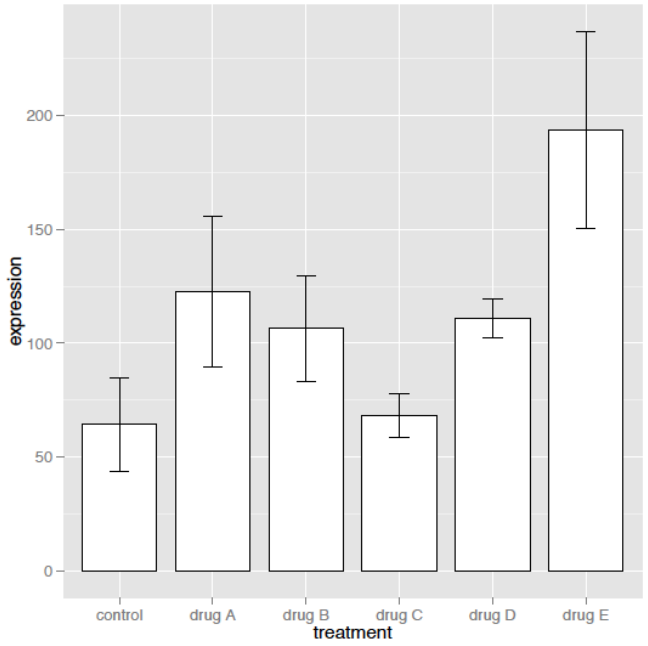
from Tobias Straub

a first test

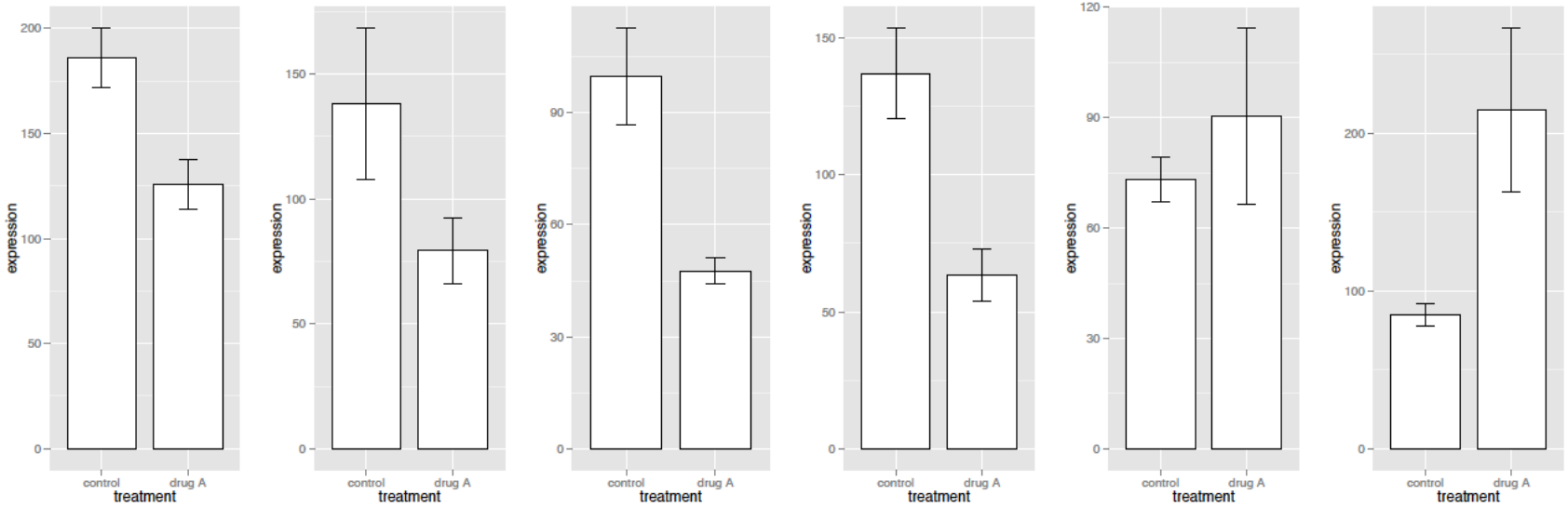


"Wow, drug A shows a significant effect, the error bars do not overlap!"

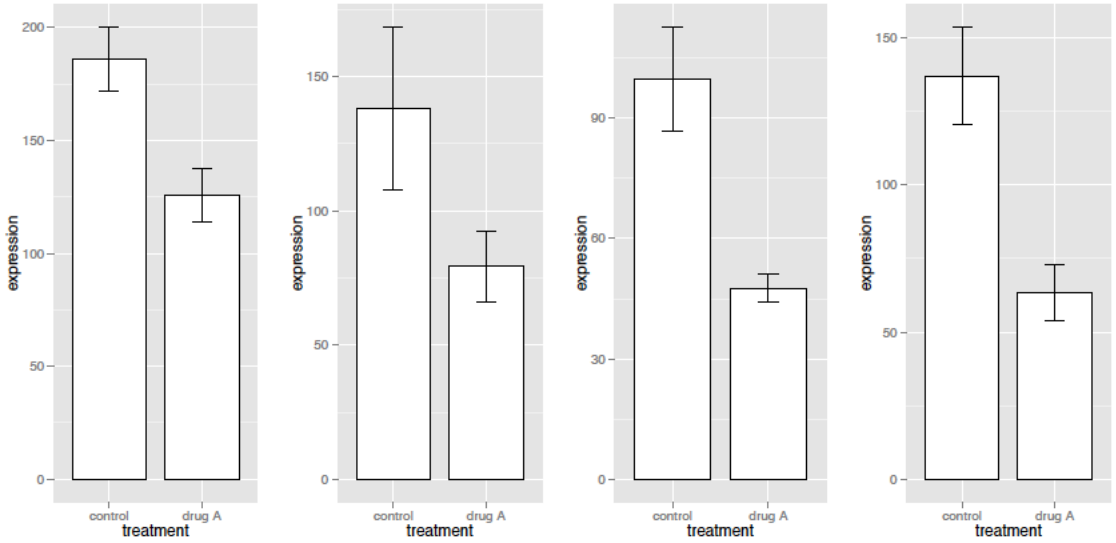
failure to repeat the result



many repetitions



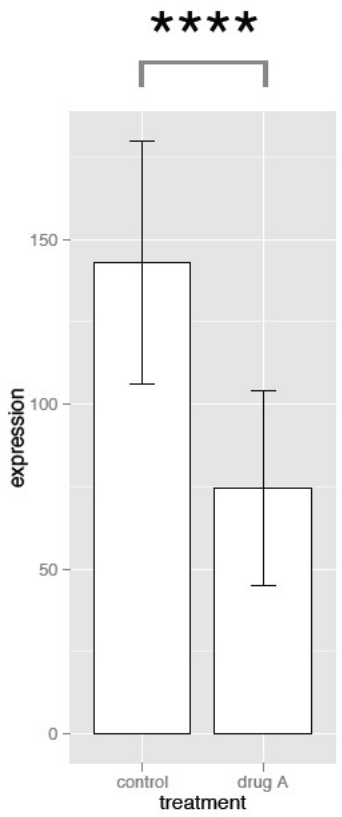
elimination of results



the statistical analysis

198.54	control
106.81	control
153.59	control
56.10	drug A
59.39	drug A
54.77	drug A
191.08	control
170.11	control
197.10	control
112.23	drug A
134.80	drug A
130.17	drug A
120.60	control
173.16	control
120.58	control
64.20	drug A
87.30	drug A
86.11	drug A
105.76	control
108.74	control
84.84	control
43.52	drug A
49.50	drug A
49.58	drug A
144.15	control
117.92	control
148.93	control
70.14	drug A
52.66	drug A
67.89	drug A

	Unpaired t test with equal SD	
1	Table Analyzed	Drug
2		
3	Column B	drug A
4	vs.	vs.
5	Column A	control
6		
7	Unpaired t test	
8	P value	< 0.0001
9	P value summary	****
10	Significantly different? (P < 0.05)	Yes
11	One- or two-tailed P value?	Two-tailed
12	t, df	t=5.592 df=28
13		
14	How big is the difference?	
15	Mean ± SEM of column A	143.8 ± 9.612, n=15
16	Mean ± SEM of column B	74.56 ± 7.642, n=15
17	Difference between means	-68.24 ± 12.20
18	95% confidence interval	-93.23 to -43.24
19	R squared	0.5276
20		
21	F test to compare variances	
22	F, DFn, Dfd	1.549, 14, 14
23	P value	0.4229
24	P value summary	ns
25	Significantly different? (P < 0.05)	No
26		



the manuscript

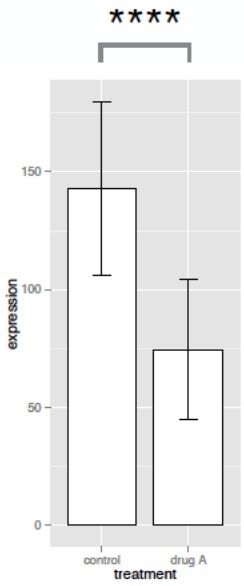


Fig.1: Drug A inhibits expression of gene x. RT-qPCR measurement of gene x transcript levels upon administration of a solvent control or 10 μM of drug A to proliferating XYZ cells for 24 hours.

Results/Discussion:
[…] we surprisingly observed an extremely significant effect of drug A on the expression of gene X […] Drug A might provide a new means to treat disease Z […]

Materials and Methods:
RT-qPCR was performed with Kit Q according Reference[1]. Statistical analysis was done in GraphPad Prism.

.. gets published, original data deleted

the manuscript

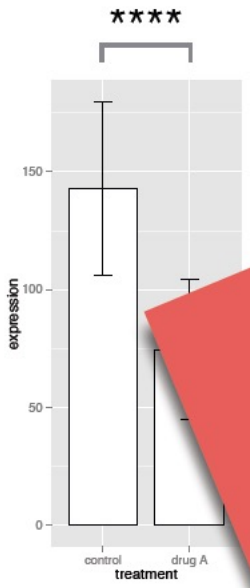


Fig.1: Drug A inhibits
expression of gene
qPCR measu
gene

Result

observed
nt effect
sion of
ht
reat

performed with Kit Q according
e[1]. Statistical analysis was done in
raphPad Prism.

irreproducible

.. gets published, original data deleted

Irreproducible because of..

- improper data presentation, interpretation and documentation
- improper treatment of replicates
- sampling bias
- improper usage of statistics
- inexistent experimental design

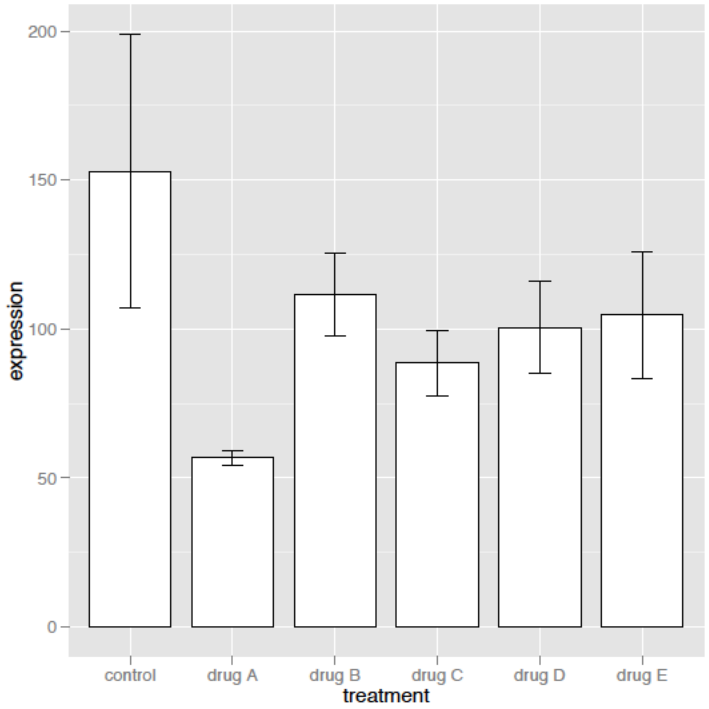
Back to the beginning

some thoughts before start pipetting

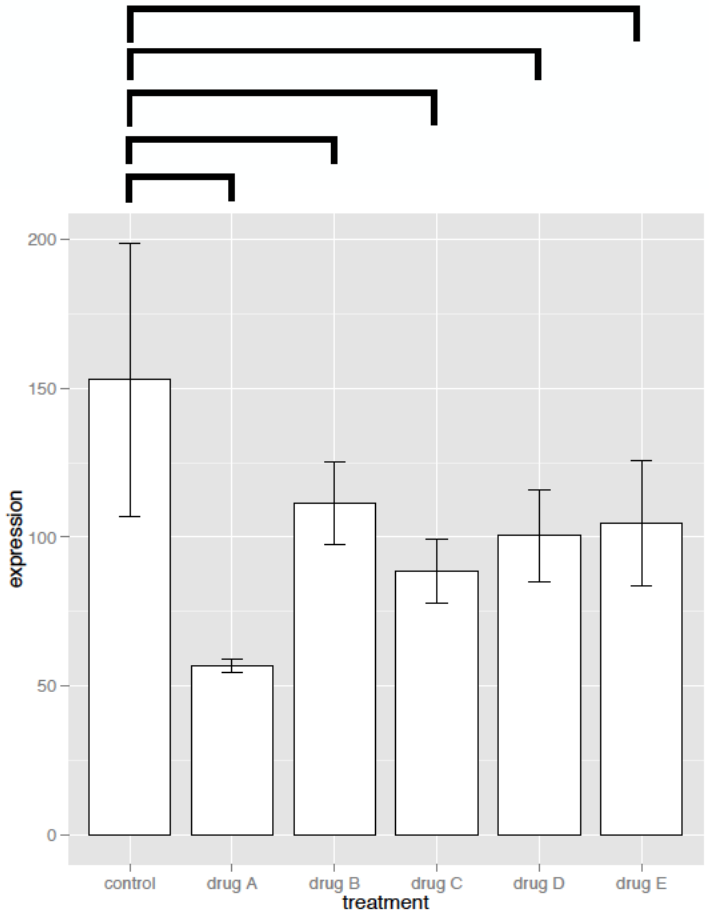
test 5 drugs on effect
on gene expression

what is the basic question?

which of the drugs (if any) has an effect on gene expression?

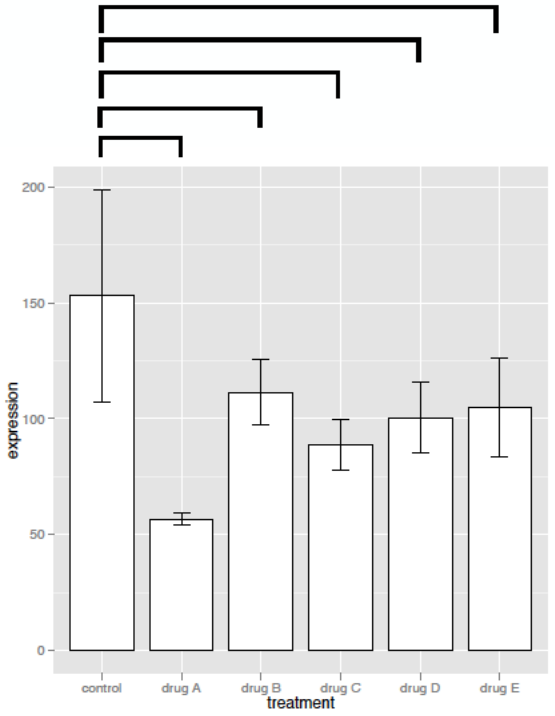


multiple testing



1-way ANOVA with Dunnett's test

- omnibus 1-way ANOVA: does any of the drugs have an effect?
- Dunnett's post test: comparing each to the control, is there an effect?



ANOVA/Dunnett requirements

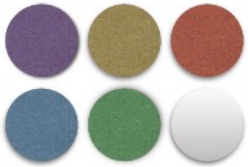
- normal distribution of data
- equal variance
- (equal group size)
- independent sampling
- representative sampling

block design, n=5, random

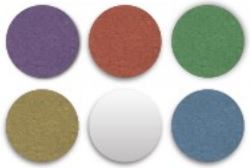
monday

tuesday

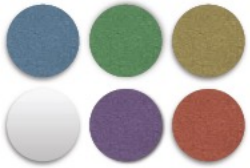
wednesday



thursday

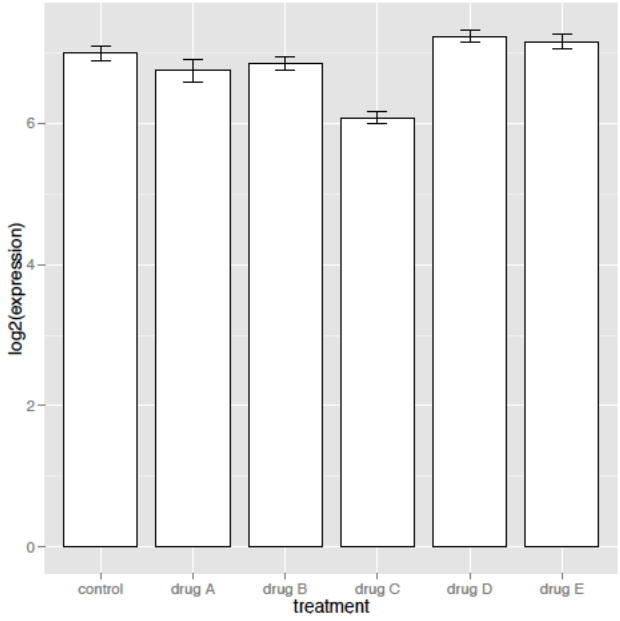


friday



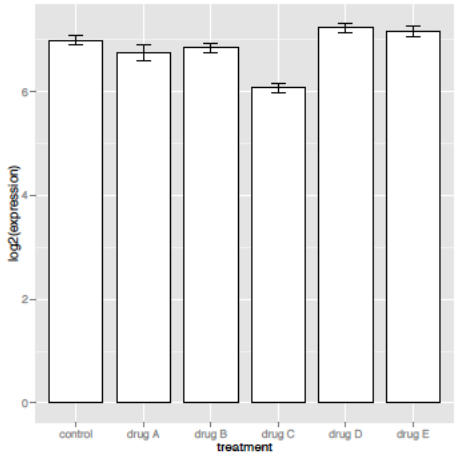
http://www.randomizer.org/form.htm

Performing the experiment



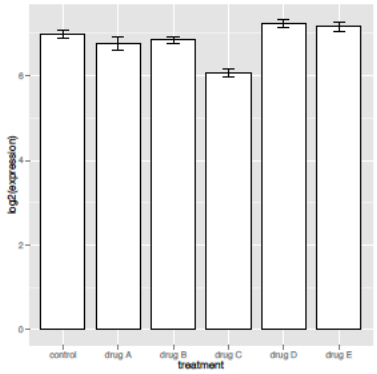
n=5,
error bars are SEM

omnibus ANOVA




```
             Df Sum Sq Mean Sq F value   Pr(>F)
treatment     5  21.74   4.348    15.2 5.62e-12 ***
Residuals   144  41.21   0.286
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dunnett's test




```
     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: aov(formula = value ~ treatment, data = ideal.measure)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
drug A - control == 0  -0.2446     0.1513  -1.617    0.352
drug B - control == 0  -0.1505     0.1513  -0.995    0.777
drug C - control == 0  -0.9158     0.1513  -6.053   <1e-04 ***
drug D - control == 0   0.2406     0.1513   1.590    0.368
drug E - control == 0   0.1649     0.1513   1.090    0.712
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

report - the figure

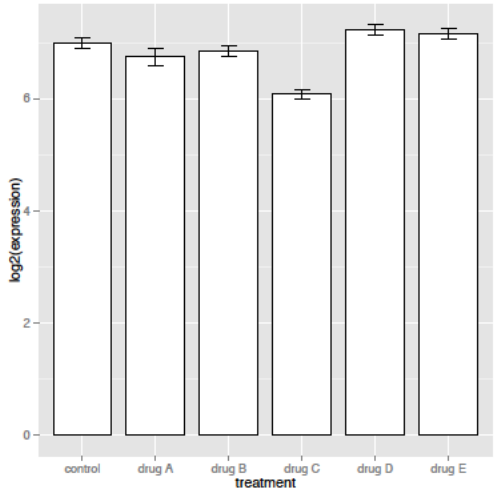


Fig.1: Drug C inhibits expression of gene x. RT-qPCR measurement of gene x transcript levels upon administration of a solvent control or 10 μM of drug A to proliferating XYZ cells for 24 hours. Error bars indicate the SEM of biological replicates (n=5).

Materials and Methods:
RT-qPCR was performed with Kit Q according to reference[1]. 5 independent biological replications were performed. Technical replicates (3 for each measurement) were averaged before analysis. Statistical analysis was done with R. 1-way ANOVA with Dunnett's post test was applied using standard parameters.

Results/Discussion:
[…] we observed changes in gene expression of gene X upon treatment with drug C (95% CI (-1.30;-0.53), p-value<0.001 (Dunnett's test)) […]

Supplementary table
1-way ANOVA and Dunnett's test result as well as raw measurement values

Towards reproducible research

- Familiarise yourself with the basic concepts of statistics and experimental design.
- Try to test simple hypotheses.
- Sample in an unbiased way.
- Keep the raw data and make it available to others.
- Report confidence intervals (of effects) and N.
- Be the most critical judge over your own data.
- Don’t trust p-values. Never.

Thanks to Tobias Straub for providing me material and ideas for this presentation.