

Applied Statistics

for Life Scientists

Tobias Straub, Biomedical Center Martinsried
tobias.straub@lmu.de

There are three kinds of lies: lies, damned lies, and statistics.
—Disraeli

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.
—H. G. Wells

EMBO Journal Checklist

1.a. How was the **sample size** chosen to ensure adequate **power** to detect a pre-specified effect size?

1.b. For animal studies, include a statement about sample size estimate even if no statistical methods were used.

2. Describe **inclusion/exclusion criteria** if samples or animals were excluded from the analysis. Were the criteria pre-established?

3. Were any steps taken to minimize the effects of subjective bias when allocating animals/samples to treatment (e.g. **randomization** procedure)? If yes, please describe.

For animal studies, include a statement about randomization even if no randomization was used.

4.a. Were any steps taken to minimize the effects of subjective bias during group allocation or/and when assessing results (e.g. **blinding** of the investigator)? If yes please describe.

4.b. For animal studies, include a statement about blinding even if no blinding was done

5. For every figure, are statistical tests justified as appropriate?

Do the data meet the assumptions of the tests (e.g., normal distribution)? Describe any methods used to assess it.

Is there an **estimate of variation within each group** of data?

Is the variance similar between the groups that are being statistically compared?



Statistics is the science of learning from data



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Create account Log in

Article Talk

Read Edit View history

Search

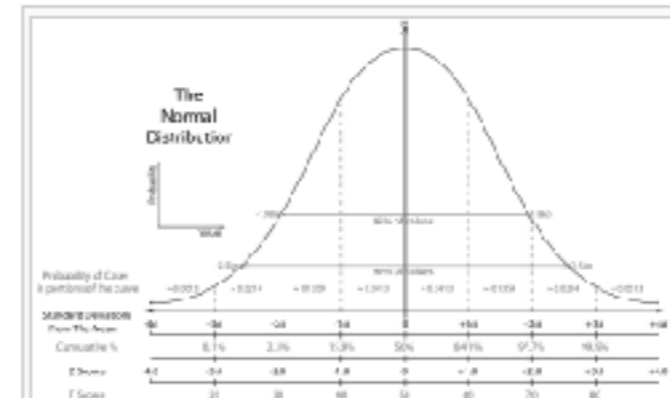
Statistics

From Wikipedia, the free encyclopedia

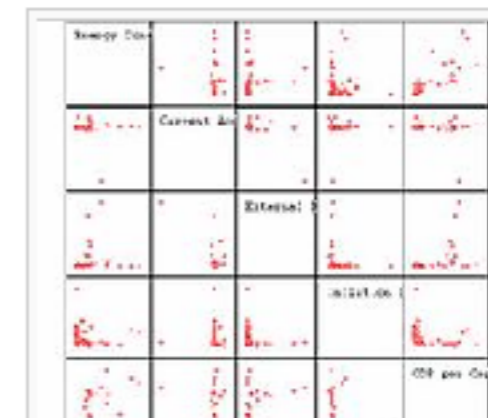
Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.^[1] In applying statistics to, e.g., a scientific, industrial, or societal problem, it is necessary to begin with a **population** or process to be studied. Populations can be diverse topics such as "all persons living in a country" or "every atom composing a crystal". It deals with all aspects of data including the planning of data collection in terms of the design of **surveys** and **experiments**.^[1]

In case **census** data cannot be collected, statisticians collect data by developing specific experiment designs and survey **samples**. Representative sampling assures that inferences and conclusions can safely extend from the sample to the population as a whole. An **experimental study** involves taking measurements of the system under study, manipulating the system, and then taking additional measurements using the same procedure to determine if the manipulation has modified the values of the measurements. In contrast, an **observational study** does not involve experimental manipulation.

Two main statistical methodologies are used in data analysis: **descriptive statistics**, which summarizes data from a sample using **indexes** such as the **mean** or **standard deviation**, and **inferential statistics**, which draws conclusions from data that are subject to random variation (e.g., observational errors, sampling variation).^[2] Descriptive statistics are most often concerned with two sets of properties of a *distribution* (sample or population): **central tendency** (or **location**)



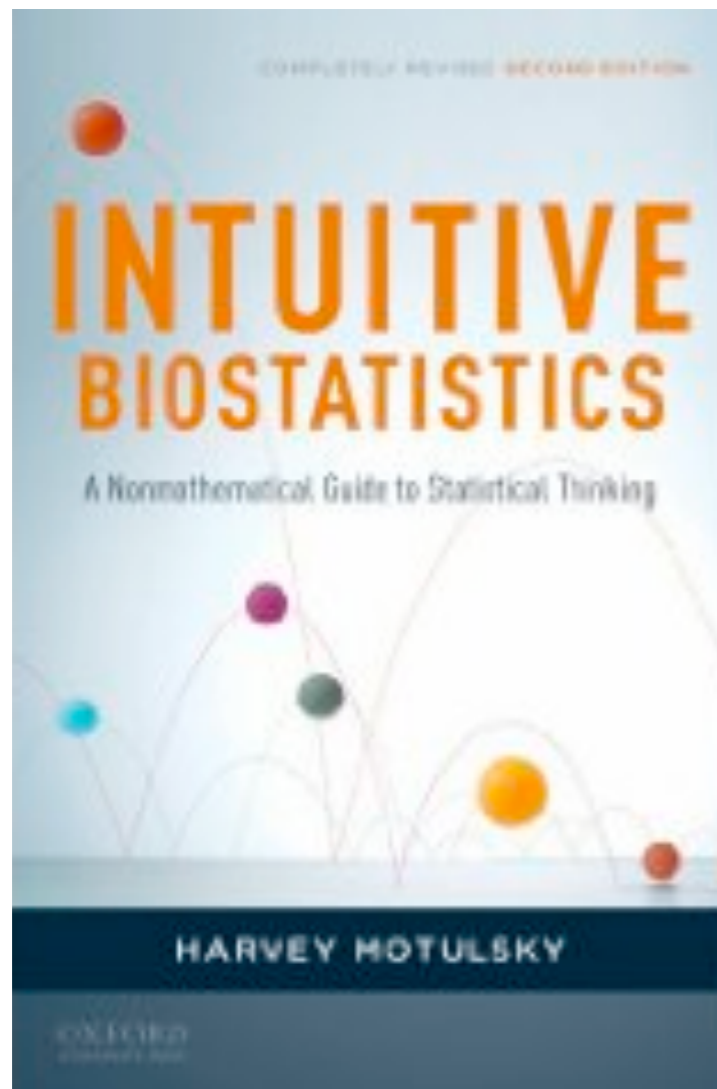
More **probability density** is found as one gets closer to the expected (mean) value in a **normal distribution**. Statistics used in **standardized testing** assessment are shown. The scales include **standard deviations**, **cumulative percentages**, **percentile equivalents**, **Z-scores**, **T-scores**, **standard nines**, and **percentages in standard nines**.



Structure

- Descriptive Statistics
- Test theory
- Common Tests
- Experimental Design / Responsible Research

Further reading



[Next topic →](#)

Basics
Introduction
Data analysis steps
Strategies of biological variables
Probability
Hypothesis testing
Random sampling
Tests for nominal variables
Exact binomial test
Fisher's exact test
Chi-square test of goodness of fit
G-test of goodness of fit
Randomization test of goodness of fit
Chi-square test of independence
G-test of independence
Fisher's exact test
Randomization test of

Introduction

Welcome to the *Handbook of Biological Statistics!* This online textbook evolved from a set of notes for my [Biological Data Analysis](#) class at the University of Delaware. My principal in that class is to teach biology students how to choose the appropriate statistical test for a particular experiment, then apply that test and interpret the results. I spend relatively little time on the mathematical basis of the tests; for most biologists, statistics is just a useful tool, like a microscope, and knowing the detailed mathematical basis of a statistical test is as unimportant to most biologists as knowing which kinds of glass were used to make a microscope lens. Biologists in very statistics-intensive fields, such as ecology, epidemiology, and systematics, may find this handbook to be a bit superficial for their needs, just as a microscopist using the latest techniques in 4-D, 3-photon confocal microscopy needs to know more about their microscope than someone who's just counting the hairs on a fly's back.

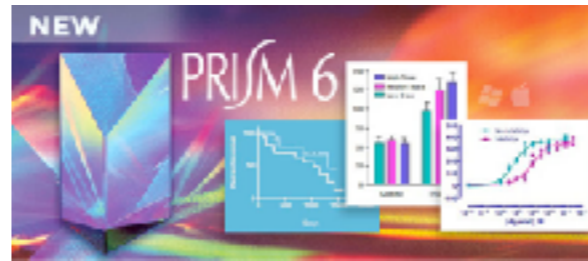
You may navigate through these pages using the "Previous topic" and "Next topic" links at the top of each page, or you may skip from topic to topic using the links on the left sidebar. Let me know if you find a broken link anywhere on these pages.

I have provided a spreadsheet to perform almost every statistical test. Each comes with sample data already entered, just download the program, replace the sample data with your data, and you'll have your answer. The spreadsheets were written for Excel, but they should also work using the free program Calc, part of the [OpenOffice.org](#) suite of programs. If you're using OpenOffice.org, some of the graphs may need re-formatting, and you may need to re-set the number of decimal places for some numbers. Let me know if you have a problem using one of the spreadsheets, and I'll try to fix it.

I've also linked to a web page for each test wherever possible. I found most of these web pages using John Pezullo's excellent [list of Interactive Statistical Calculation Pages](#), which is a good place

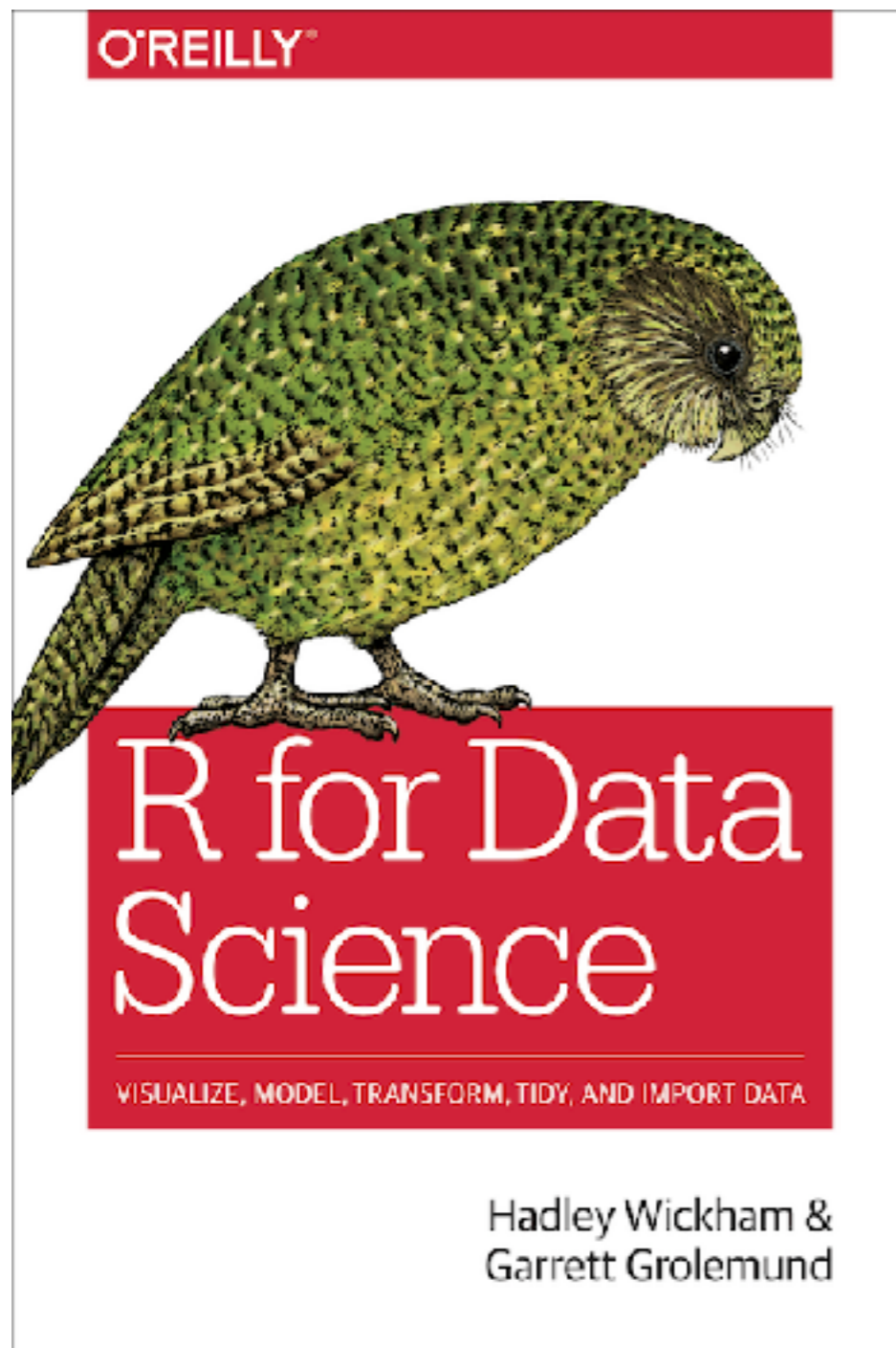
[http://udel.edu/~mcdonald/
statintro.html](http://udel.edu/~mcdonald/statintro.html)

Software

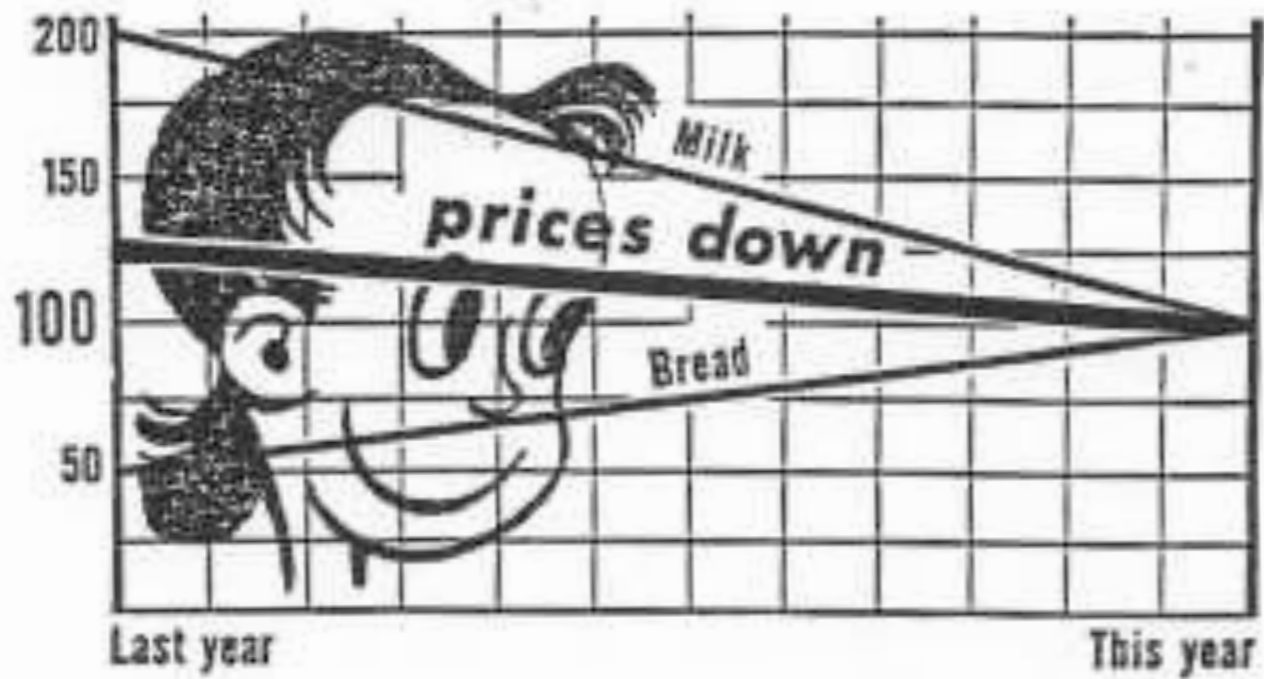
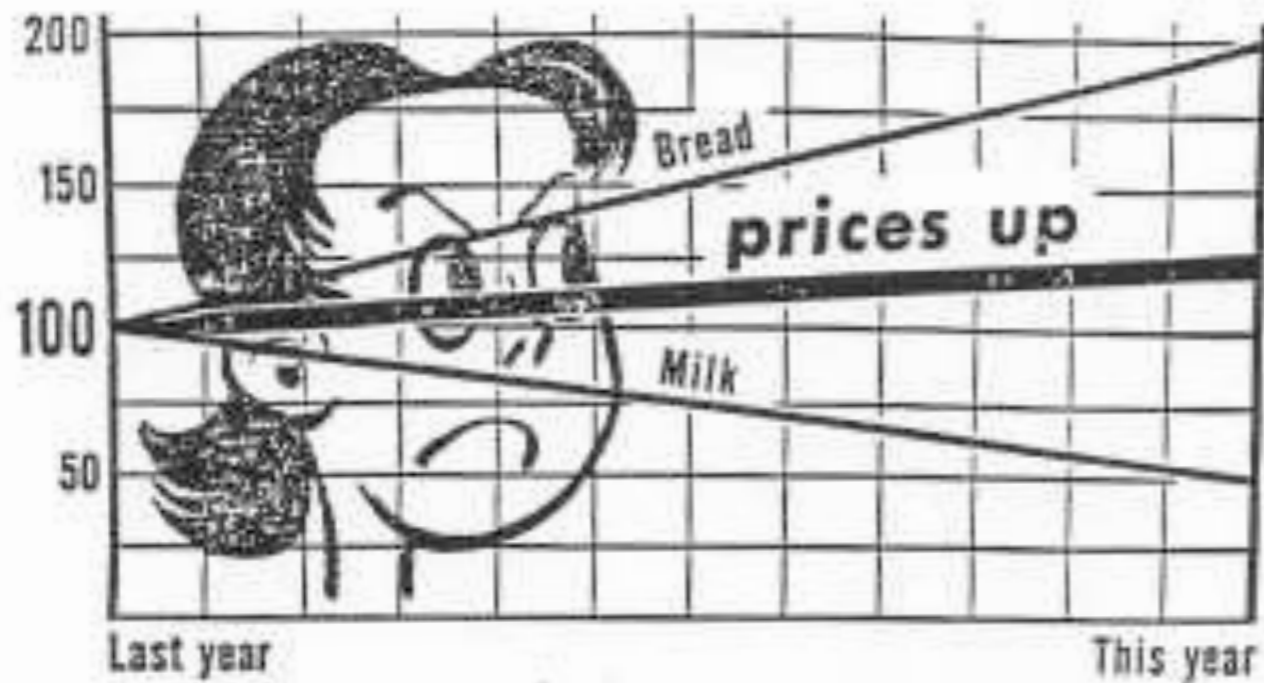


	EXCEL	Prism	R
price	medium	high	free
ease of use	medium	easy	difficult
coverage	low	high	infinite
misusage	average	made easy	average
graphs	poor/limited	good	best/flexible
other	NOT a stats app		huge community de facto standard

Learning R

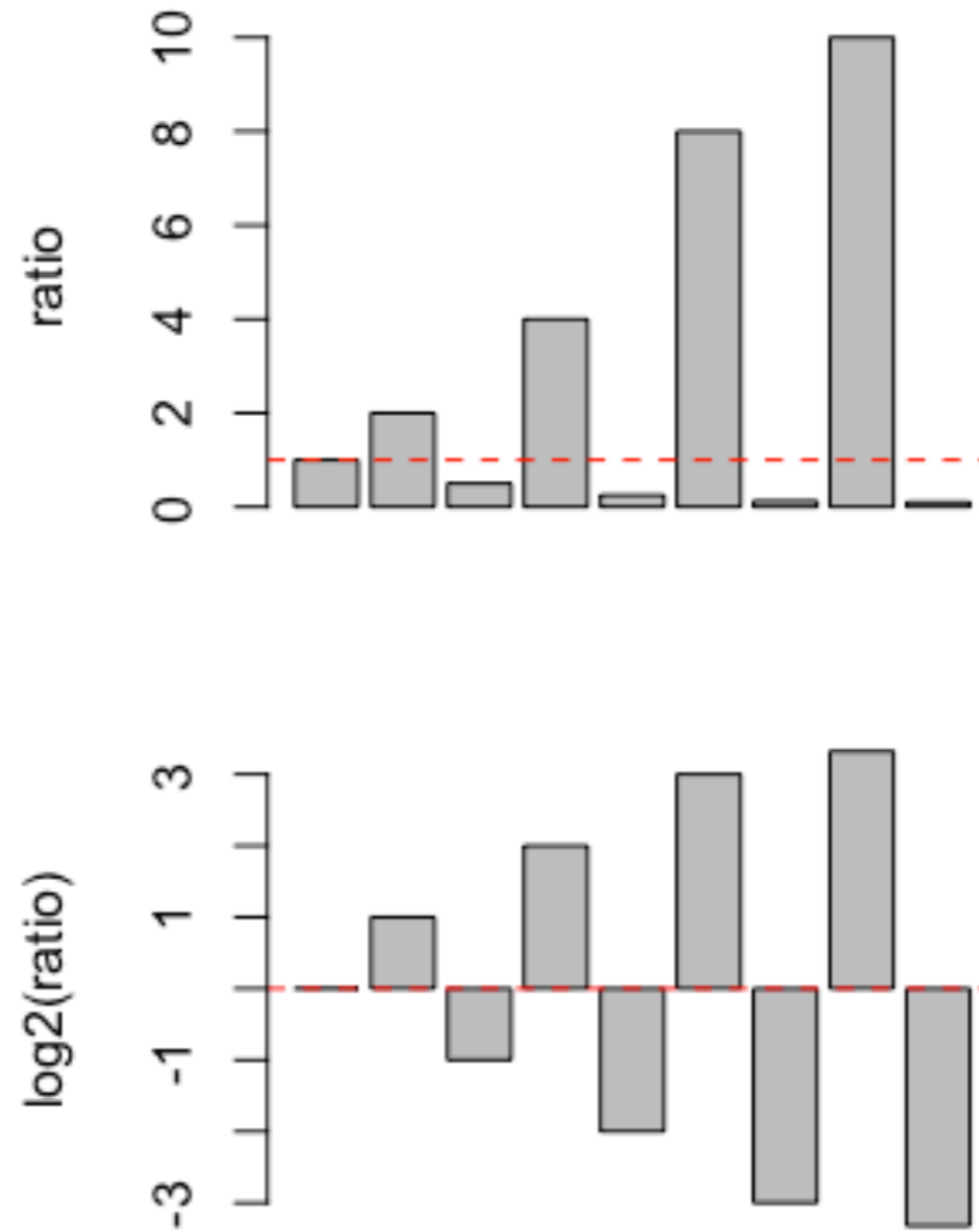


<https://r4ds.had.co.nz>

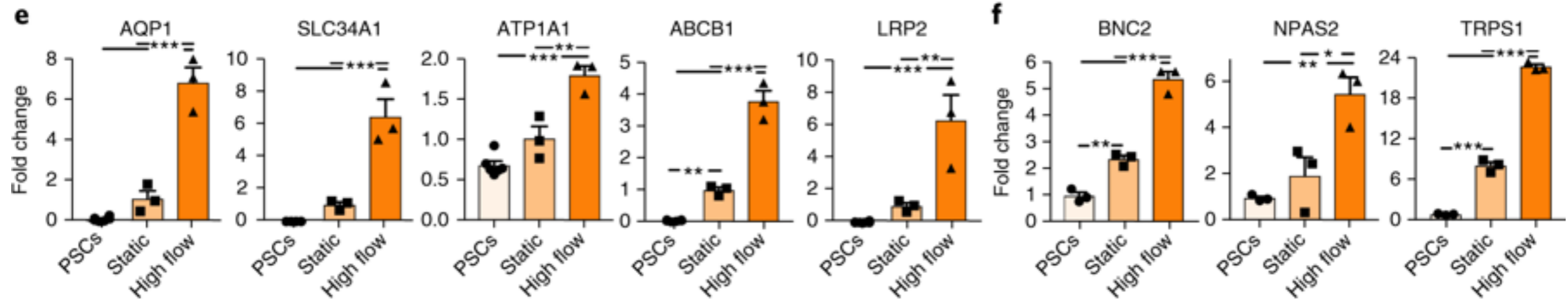


Descriptive Statistics

Ratios in linear versus log space



Ratios are not the only problem here..



d–f represent three technical replicates on RNA pooled from 6 organoids (biological replicates) per condition. Statistical analysis for **d–f,h,i** was determined at a value of $P < 0.05$ as determined by one-way ANOVA with Tukey's multiple-comparisons test. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Data, what is it?

cases
samples
observations

variables

values

	A	B	C	D	E	F	G	H
1	Pregnancies	Glucose	Blood Pressure	Skin	BMI	Diabetes Pedigree Function	Age	Diabetic
2	6	148	72	35	33.6	0.627	50	Yes
3	1	85	66	29	26.6	0.351	31	No
4	1	89	66	23	28.1	0.167	21	No
5	3	78	50	32	31	0.248	26	Yes
6	2	197	70	45	30.5	0.158	53	Yes
7	5	166	72	19	25.8	0.587	51	Yes
8	0	118	84	47	45.8	0.551	31	Yes
9	1	103	30	38	43.3	0.183	33	No
10	3	126	88	41	39.3	0.704	27	No
11	9	119	80	35	29	0.263	29	Yes
12	1	97	66	15	23.2	0.487	22	No
13	5	109	75	26	36	0.546	60	No
14	3	88	58	11	24.8	0.267	22	No
15	10	122	78	31	27.6	0.512	45	No
16	4	103	60	33	24	0.966	33	No
17	9	102	76	37	32.9	0.665	46	Yes
18	2	90	68	42	38.2	0.503	27	Yes
19	4	111	72	47	37.1	1.39	56	Yes
20	3	180	64	25	34	0.271	26	No
21	7	106	92	18	22.7	0.235	48	No
22	9	171	110	24	45.4	0.721	54	Yes
23	0	180	66	39	42	1.893	25	Yes
24	2	71	70	27	28	0.586	22	No
25	1	103	80	11	19.4	0.491	22	No

a collection of measurements of similar structure



The American Statistician



ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: <https://www.tandfonline.com/loi/utas20>

Data Organization in Spreadsheets

Karl W. Broman & Kara H. Woo

To cite this article: Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: [10.1080/00031305.2017.1375989](https://doi.org/10.1080/00031305.2017.1375989)

To link to this article: <https://doi.org/10.1080/00031305.2017.1375989>

Best of Data Organisation in Spread Sheets

- Be consistent
- Choose Good Names for Things
- Put Just One Thing in a Cell
- No Empty Cells
- Make it a Rectangle
- No Calculations in the Raw Data Files
- Do Not Use Font Color or Highlighting as Data
- **Do_not_use_white_space_but_underscores_for_names**

the origin of data matters.. a lot

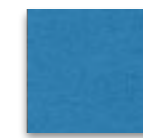
- observational (descriptive) or experimental (controlled)?
- sampling strategy
- Metadata (what, when, who, how) matters

Data types

- **Continuous** data
numerical data which can hold any value
- **Discrete** data
numerical data which can only take certain values
- **Categorical** data
Variables are labels of grouped features
(classifications)

Data Types

	A	B	C	D	E	F	G	H
1	Pregnancies	Glucose	Blood Pressure	Skin	BMI	Diabetes Pedigree Function	Age	Diabetic
2	6	148	72	35	33.6	0.627	50	Yes
3	1	85	66	29	26.6	0.351	31	No
4	1	89	66	23	28.1	0.167	21	No
5	3	78	50	32	31	0.248	26	Yes
6	2	197	70	45	30.5	0.158	53	Yes
7	5	166	72	19	25.8	0.587	51	Yes
8	0	118	84	47	45.8	0.551	31	Yes
9	1	103	30	38	43.3	0.183	33	No
10	3	126	88	41	39.3	0.704	27	No
11	9	119	80	35	29	0.263	29	Yes
12	1	97	66	15	23.2	0.487	22	No
13	5	109	75	26	36	0.546	60	No
14	3	88	58	11	24.8	0.267	22	No
15	10	122	78	31	27.6	0.512	45	No
16	4	103	60	33	24	0.966	33	No
17	9	102	76	37	32.9	0.665	46	Yes
18	2	90	68	42	38.2	0.503	27	Yes
19	4	111	72	47	37.1	1.39	56	Yes
20	3	180	64	25	34	0.271	26	No
21	7	106	92	18	22.7	0.235	48	No
22	9	171	110	24	45.4	0.721	54	Yes
23	0	180	66	39	42	1.893	25	Yes
24	2	71	70	27	28	0.586	22	No
25	1	103	80	11	19.4	0.491	22	No



discrete

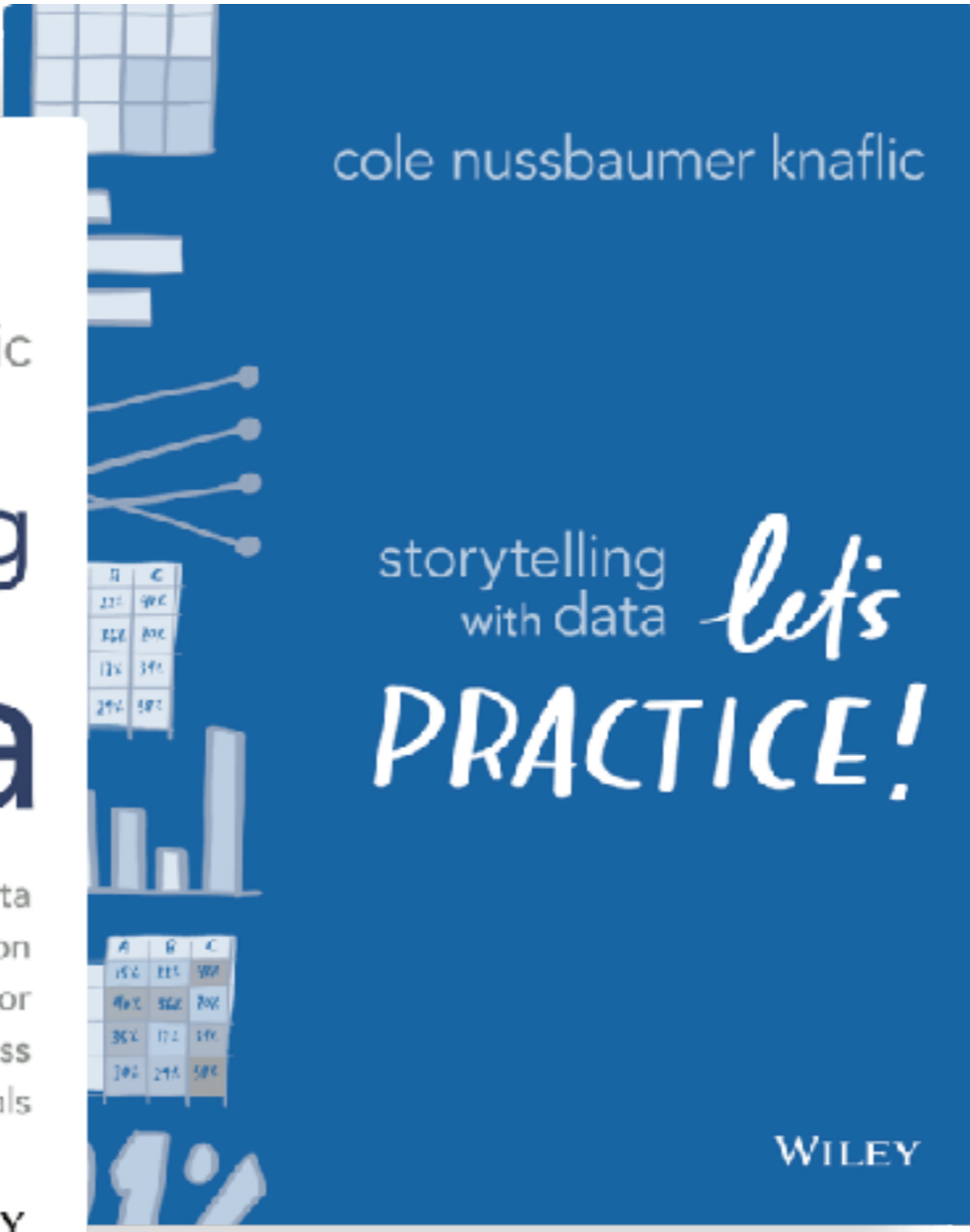
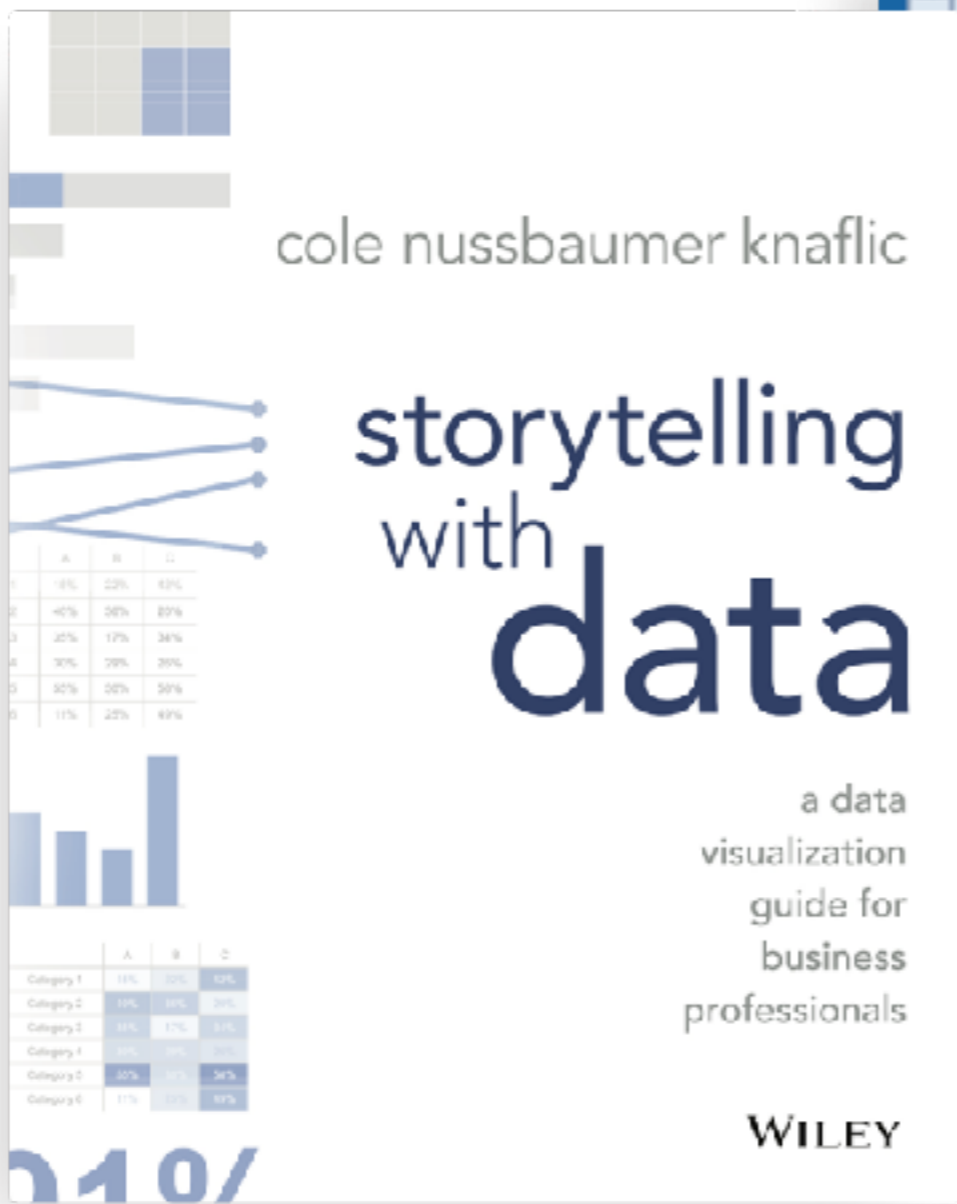


continuous



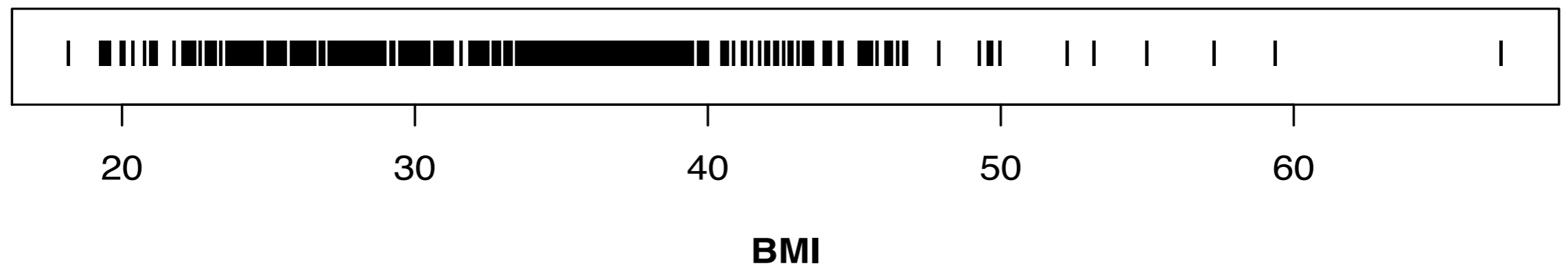
categorical

visual representation



Plotting all data points (continuous data)

Example:
BMI of 532 Pima Indian Females



stripchart

more (too many) data points

```
expr.value
1616608_a_at 9.118380
1622892_s_at 8.115987
1622893_at 2.194861
1622894_at 2.194861
1622895_at 8.871565
1622896_at 8.762262
1622897_at 2.194861
1622898_a_at 9.422677
1622899_at 3.987549
1622900_at 2.194861
1622901_at 2.194861
1622902_at 2.195272
1622903_s_at 7.679026
1622904_at 2.212932
1622905_at 2.203904
1622906_at 2.198816
1622907_at 8.294115
1622908_a_at 11.002117
1622909_at 10.899726
1622910_at 2.194861
1622911_at 2.194861
1622912_at 7.421109
1622913_a_at 2.194861
1622914_at 2.194861
1622915_at 2.194861
1622916_at 2.274991
1622917_a_at 2.194861
1622918_at 2.289296
1622919_at 2.195047
1622920_at 3.757421
```

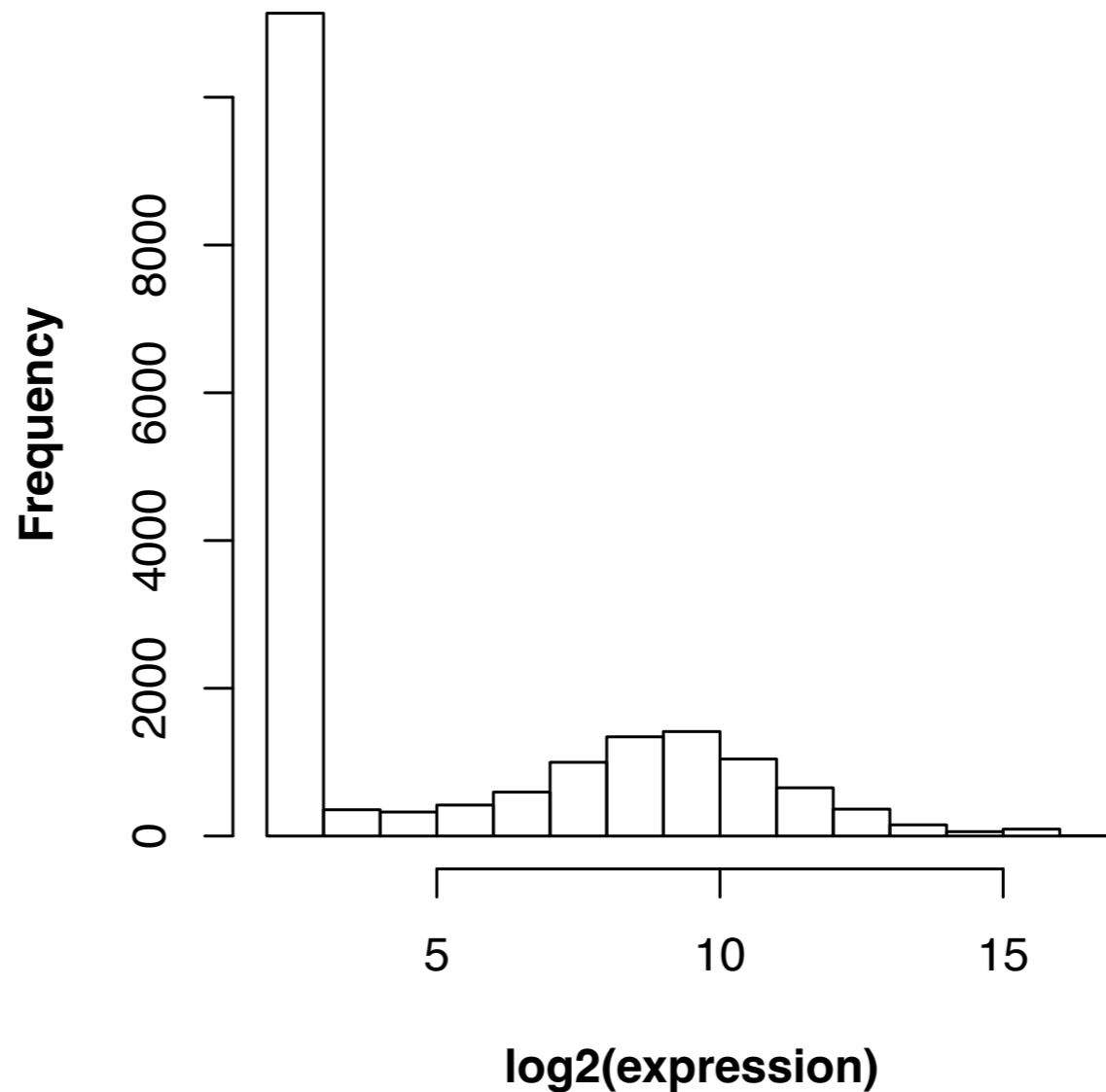
...

n = 18952

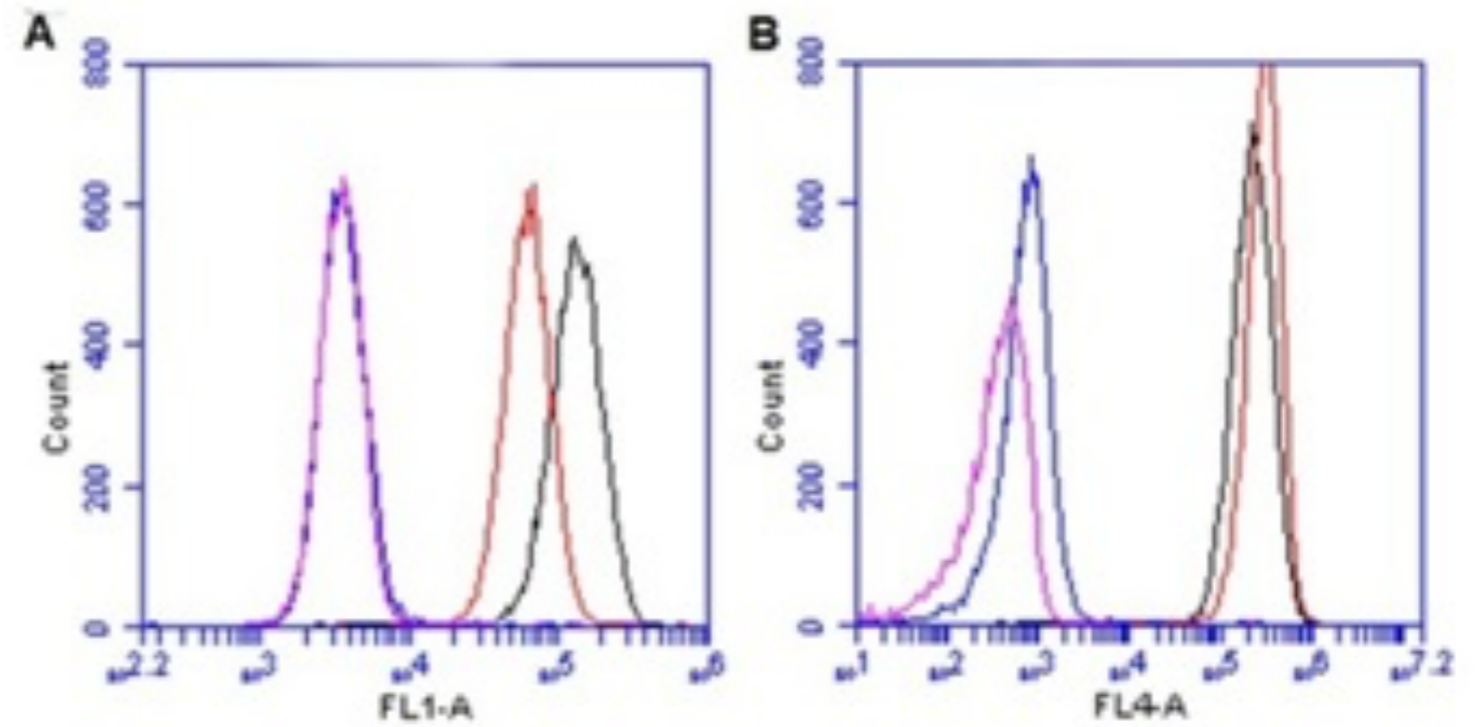
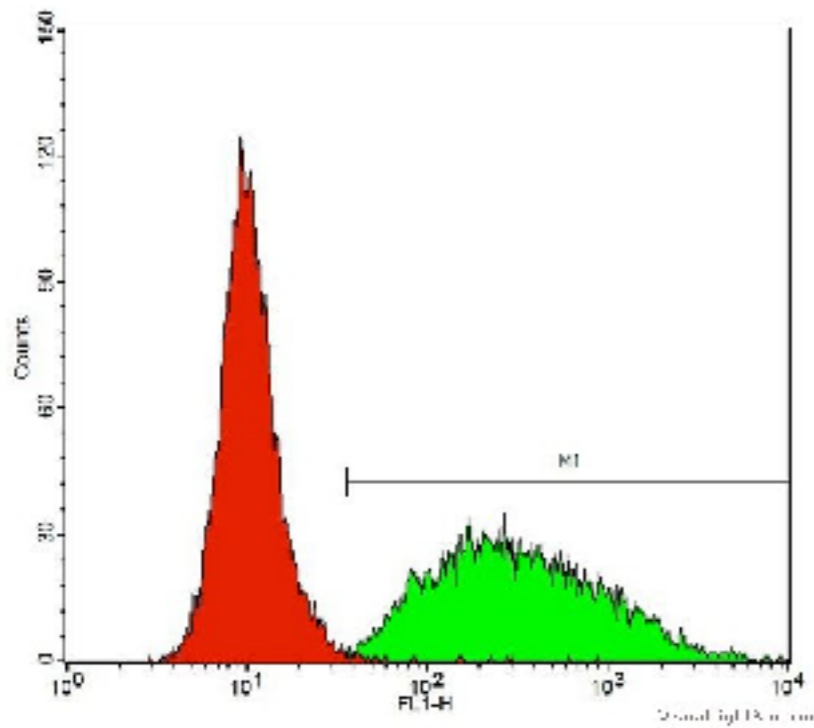
Continuous Variables - Histogram

	expr.value
1616608_a_at	9.118380
1622892_s_at	8.115987
1622893_at	2.194861
1622894_at	2.194861
1622895_at	8.871565
1622896_at	8.762262
1622897_at	2.194861
1622898_a_at	9.422677
1622899_at	3.987549
1622900_at	2.194861
1622901_at	2.194861
1622902_at	2.195272
1622903_s_at	7.679026
1622904_at	2.212932
1622905_at	2.203904
1622906_at	2.198816
1622907_at	8.294115
1622908_a_at	11.002117
1622909_at	10.899726
1622910_at	2.194861
1622911_at	2.194861
1622912_at	7.421109
1622913_a_at	2.194861
1622914_at	2.194861
1622915_at	2.194861
1622916_at	2.274991
1622917_a_at	2.194861
1622918_at	2.289296
1622919_at	2.195047
1622920_at	3.757421

...

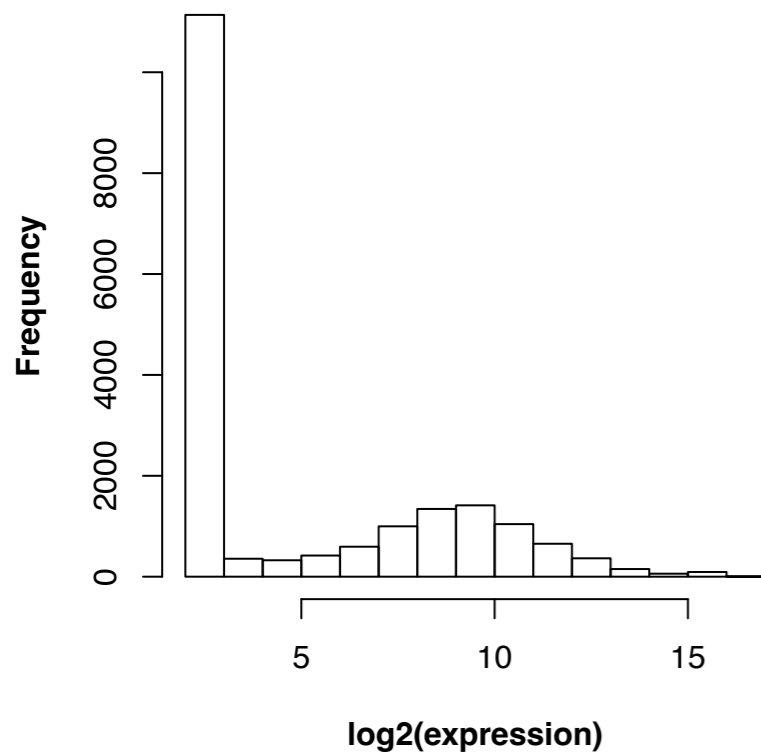


Histogram - Flow Cytometry

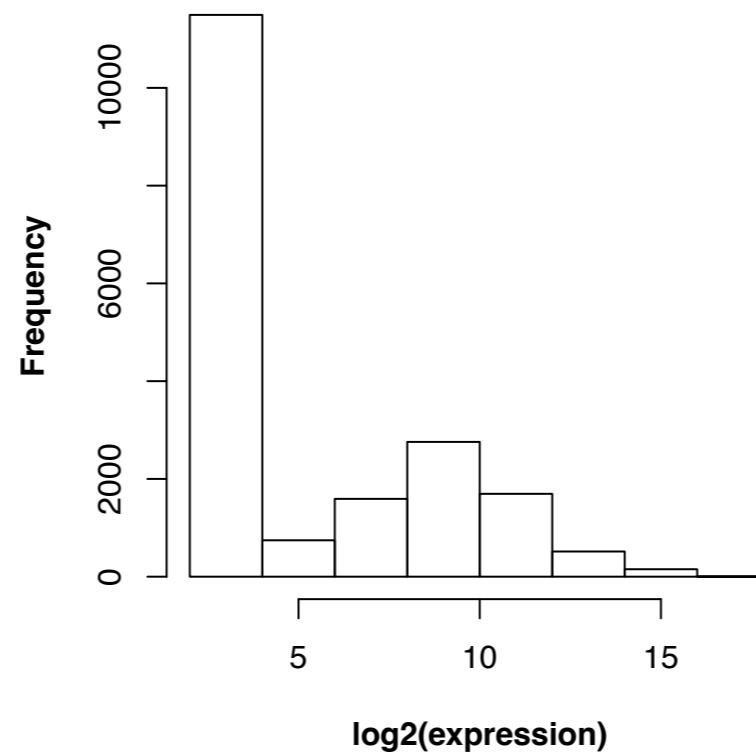


Continuous Variables - Histogram

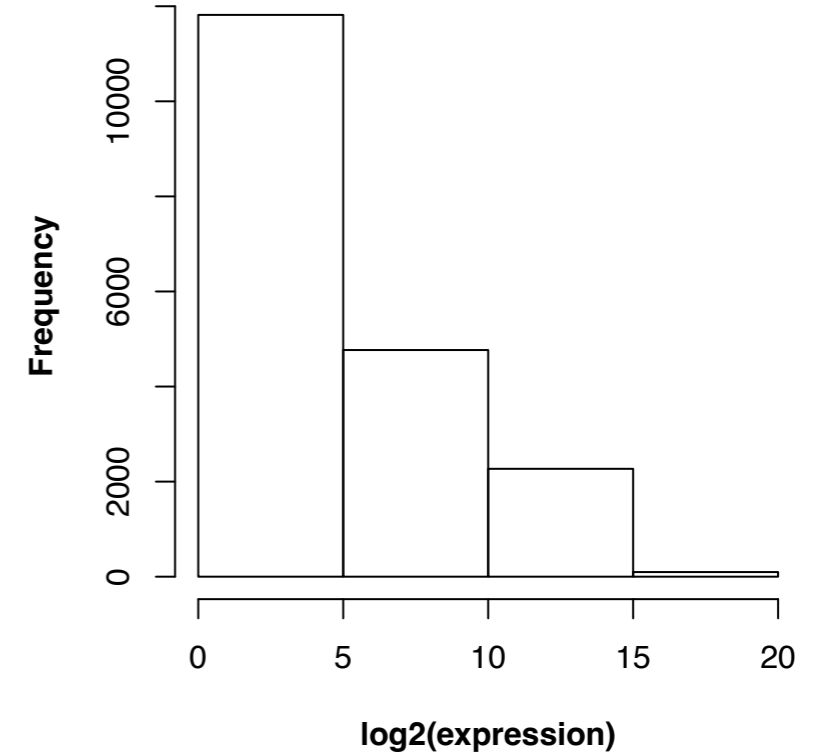
15



10

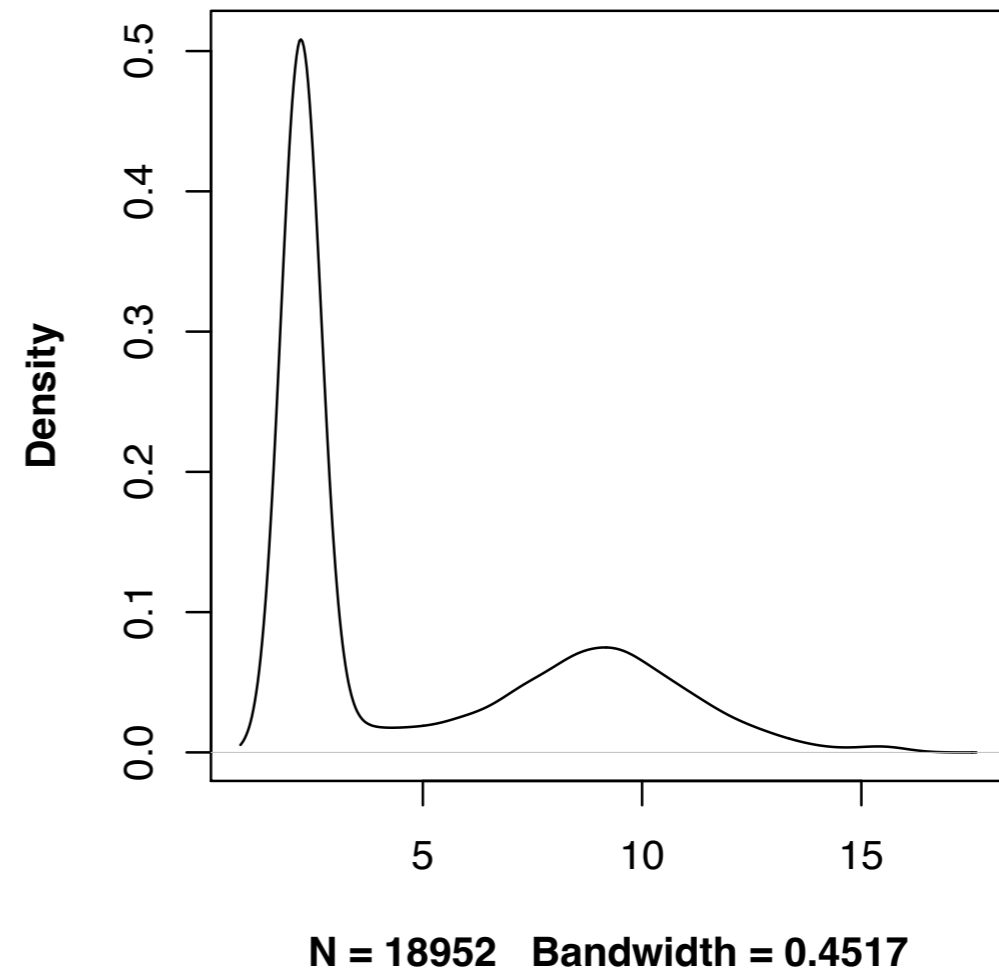
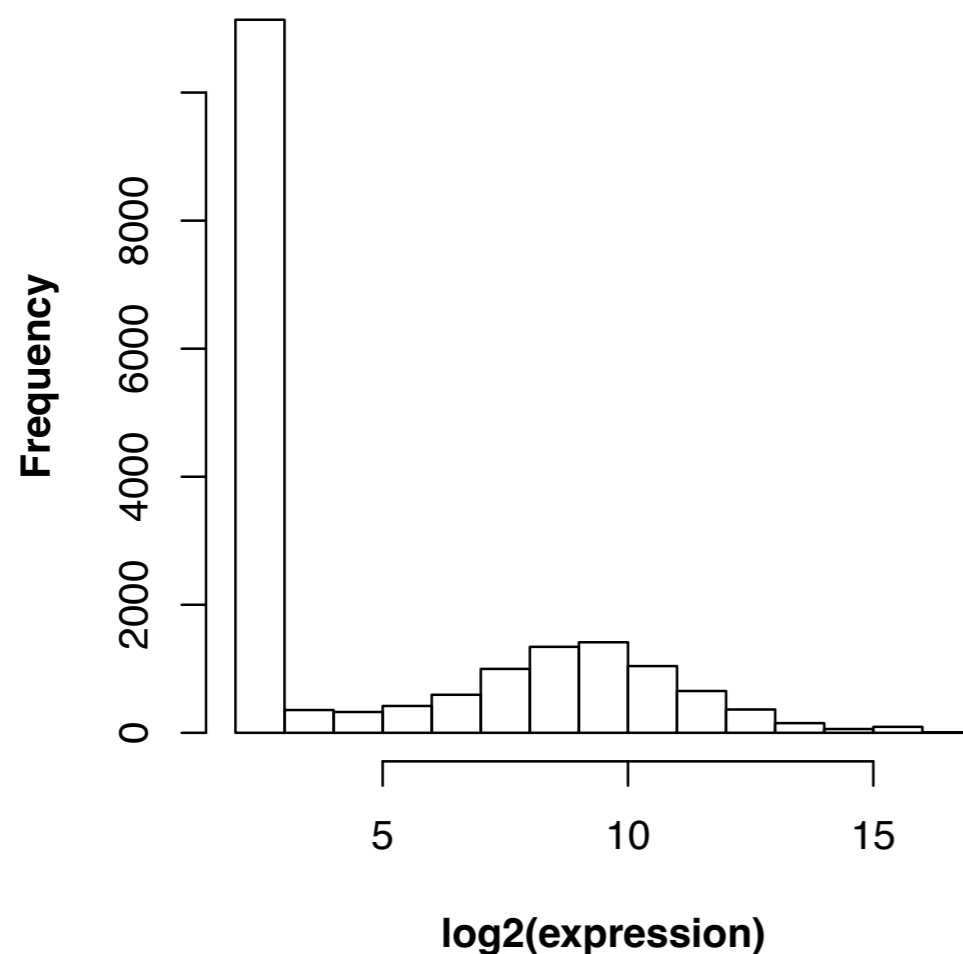


5



The size of the bins (= width of the bars) is a matter of choice and has to be determined sensibly!

Continuous Variables - Density Plot

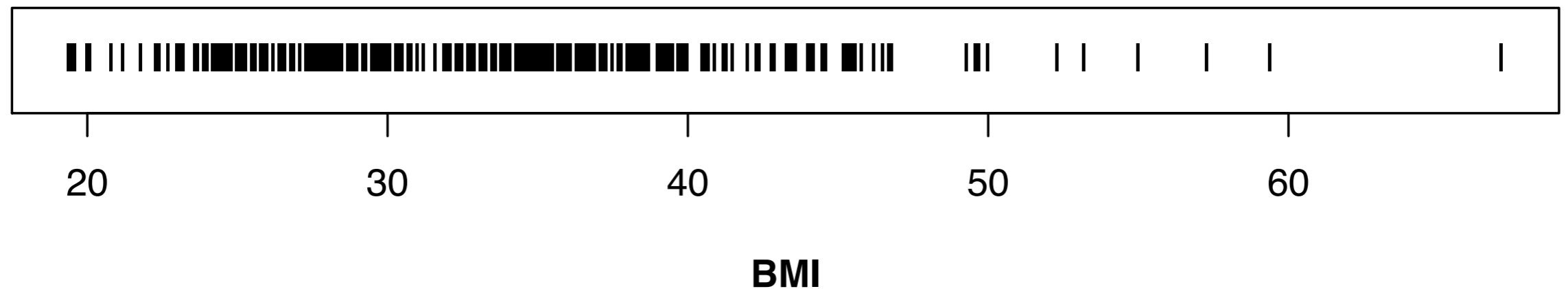


Data will be smoothed automatically.
This is very suggestive and blurs discontinuities in a distribution

non-visual description

Measures of Location and Scatter

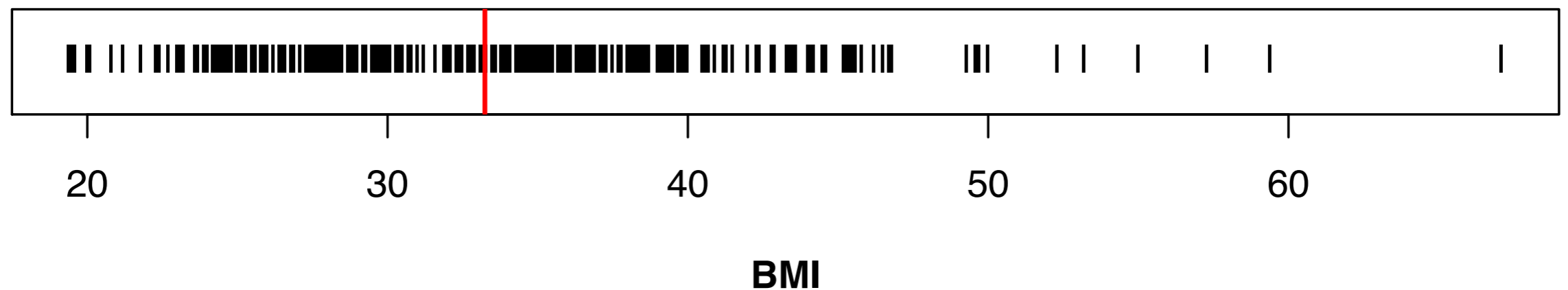
Example:
BMI of 332 Pima Indian Females



Measures of Location and Scatter

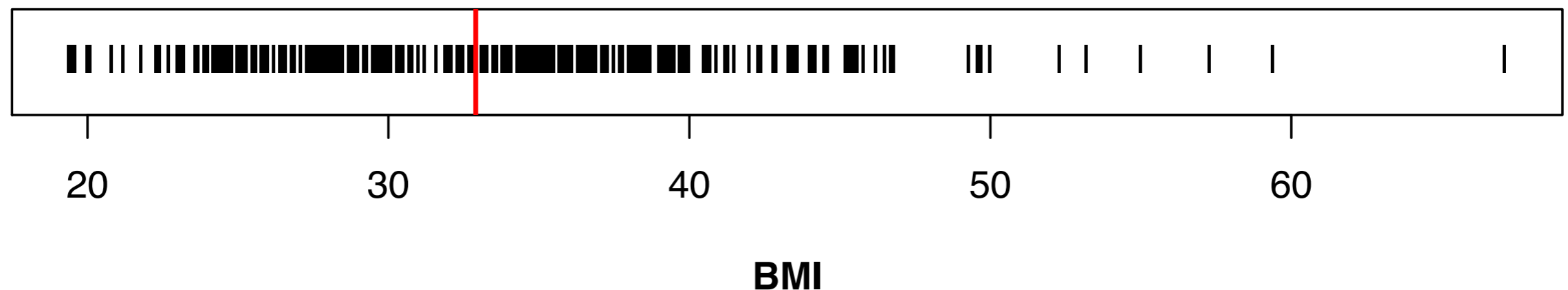
Mean:

sum of all observations/number of samples



Measures of Location and Scatter

Median:
a number M
such that 50% of all observations
are less than or equal to M ,
and 50% are greater than or equal to M



Mean vs. Median

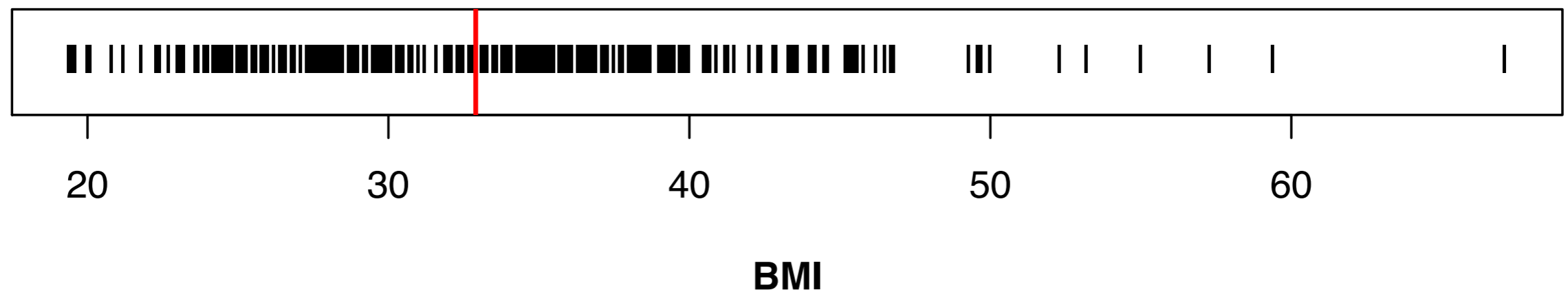
- **median** should be preferred to the **mean** if the value distribution
 - a) is asymmetric
 - b) has extreme outliers
- the **mean** is more precise than the **median** if the distribution is approximately normal

Continuous Variables - Quantiles

Quantile:

The p -quantile is a property value that splits a distribution. On the left of the p -quantile are $100 * p$ percent of all values. On the right are $100 * (1 - p)$ percent of all values.

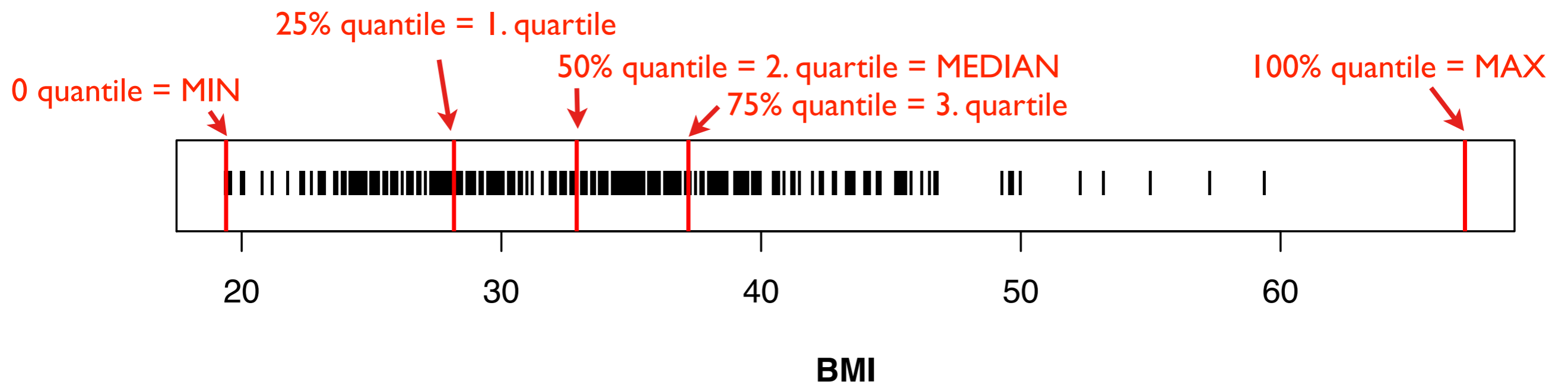
50% quantile = MEDIAN



Continuous Variables - Quantiles

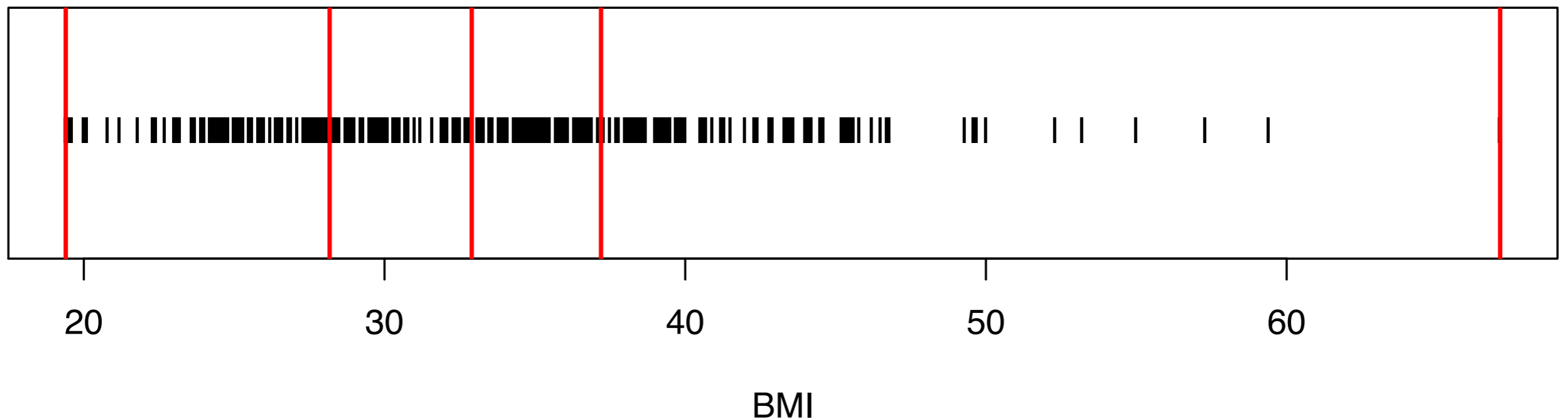
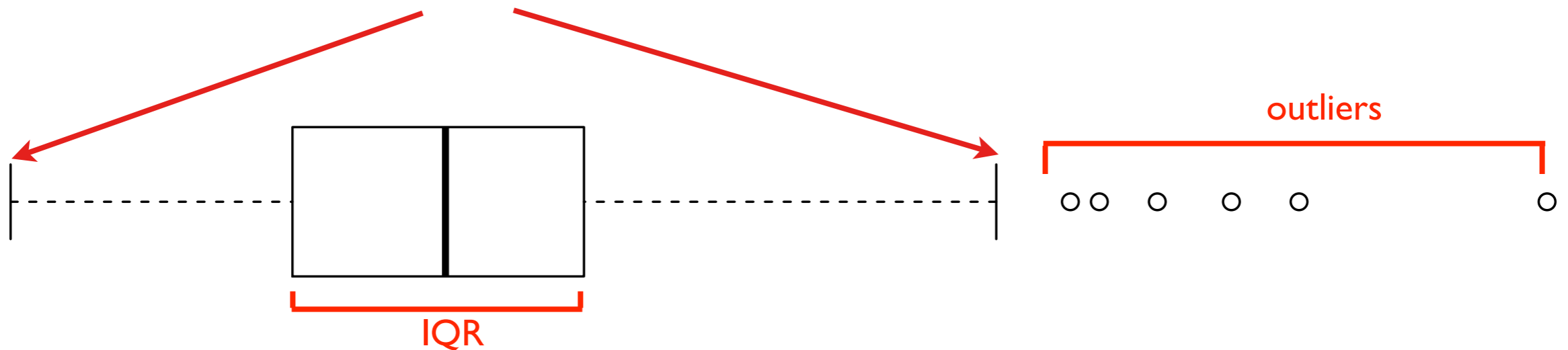
Quantile:

The p -quantile is a property value that splits a distribution. On the left of the p -quantile are $100 * p$ percent of all values. On the right are $100 * (1 - p)$ percent of all values.

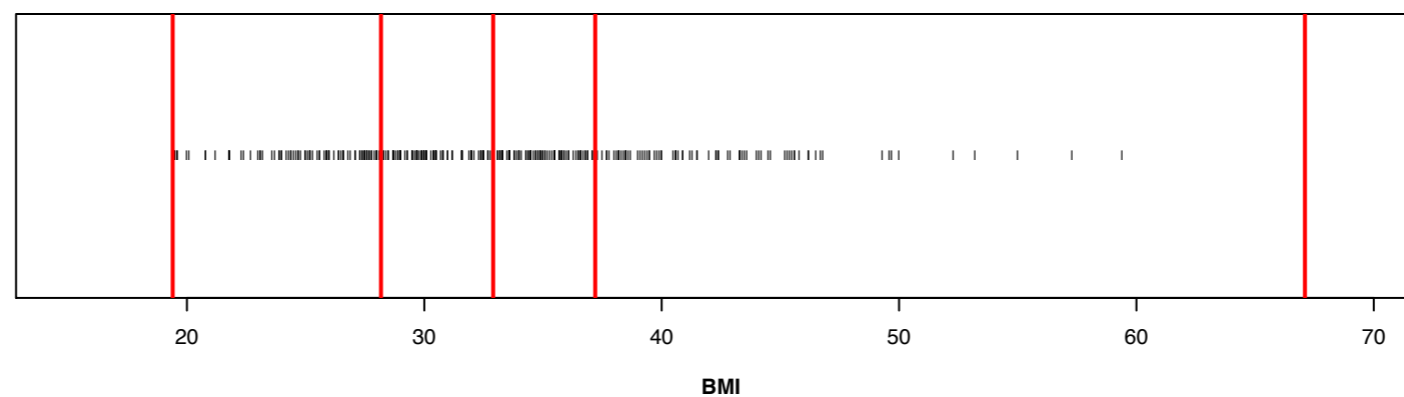
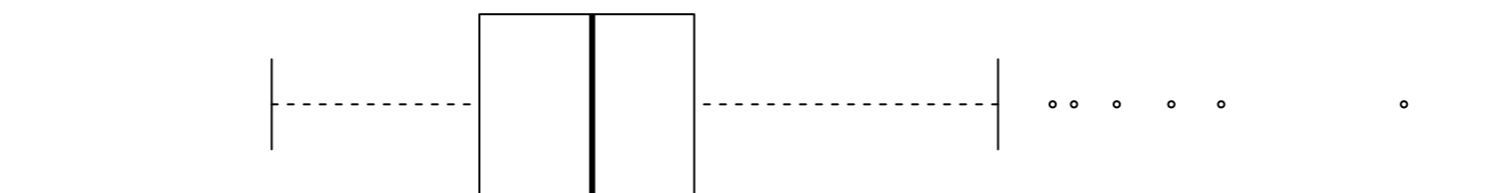
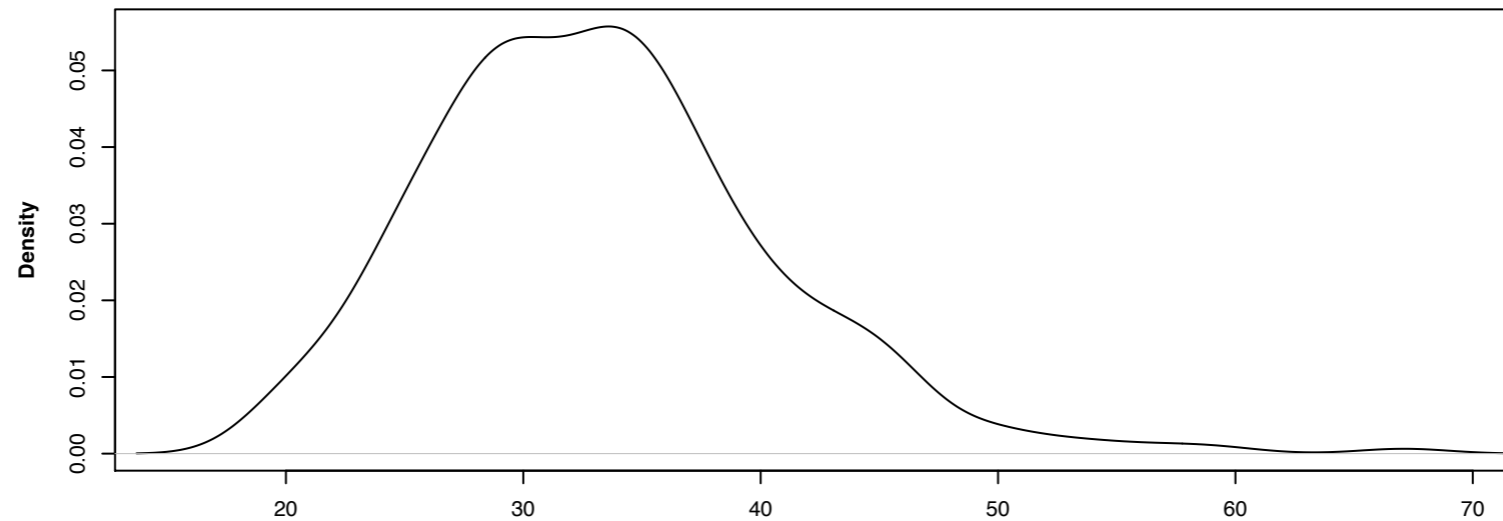


Continuous Variables - Boxplot

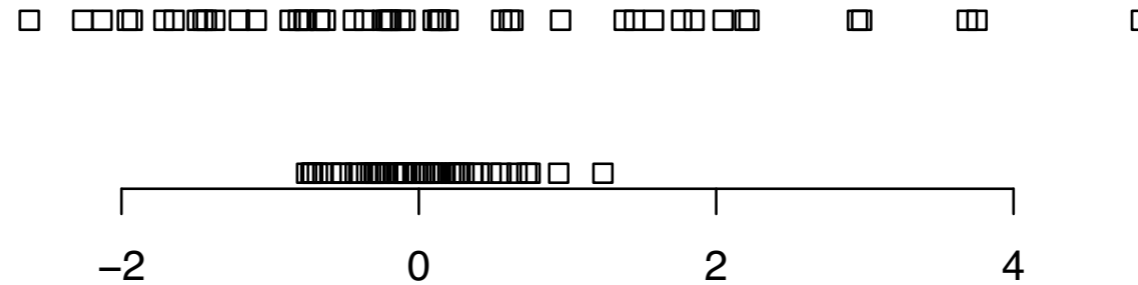
Whisker
maximally $1.5 \times$ interquartile distance (IQR),
ends at the last data point falling within this
range or min/max



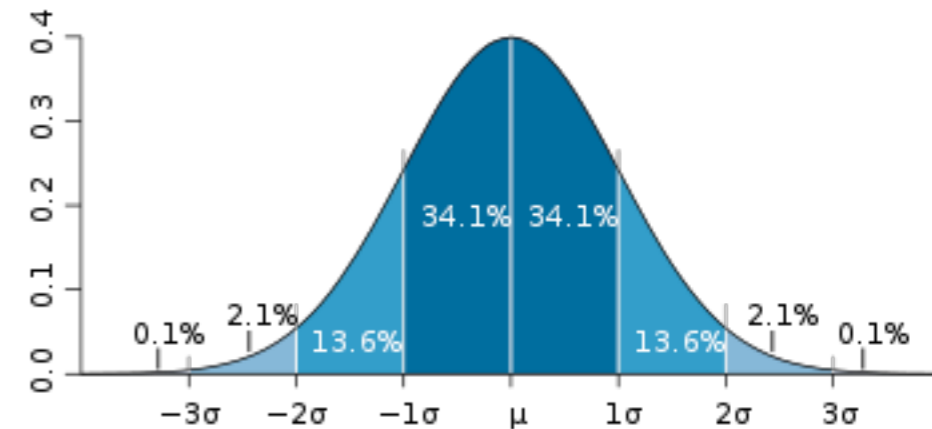
Visual continuous data representation

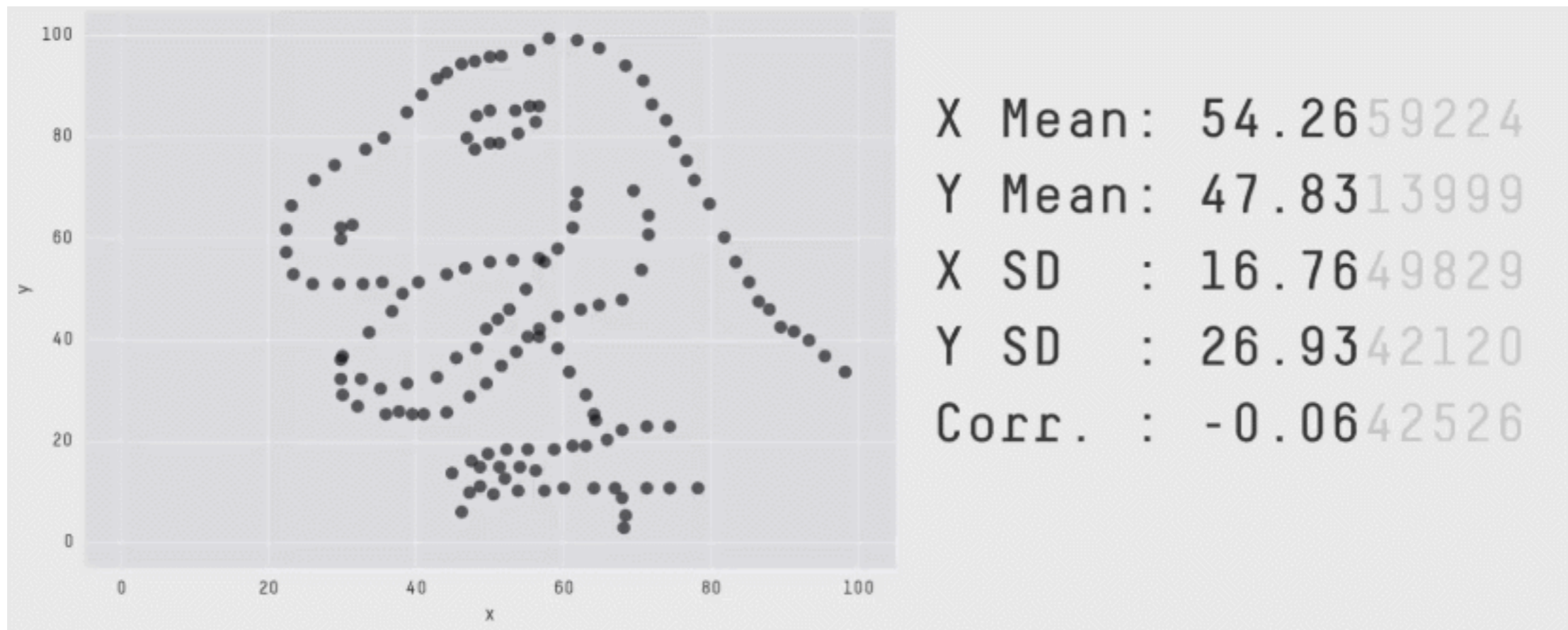


Description of Scatter



- variance
= mean squared deviation of mean
- standard deviation
= square root of the variance
- IQR





<https://www.autodeskresearch.com/publications/samestats>

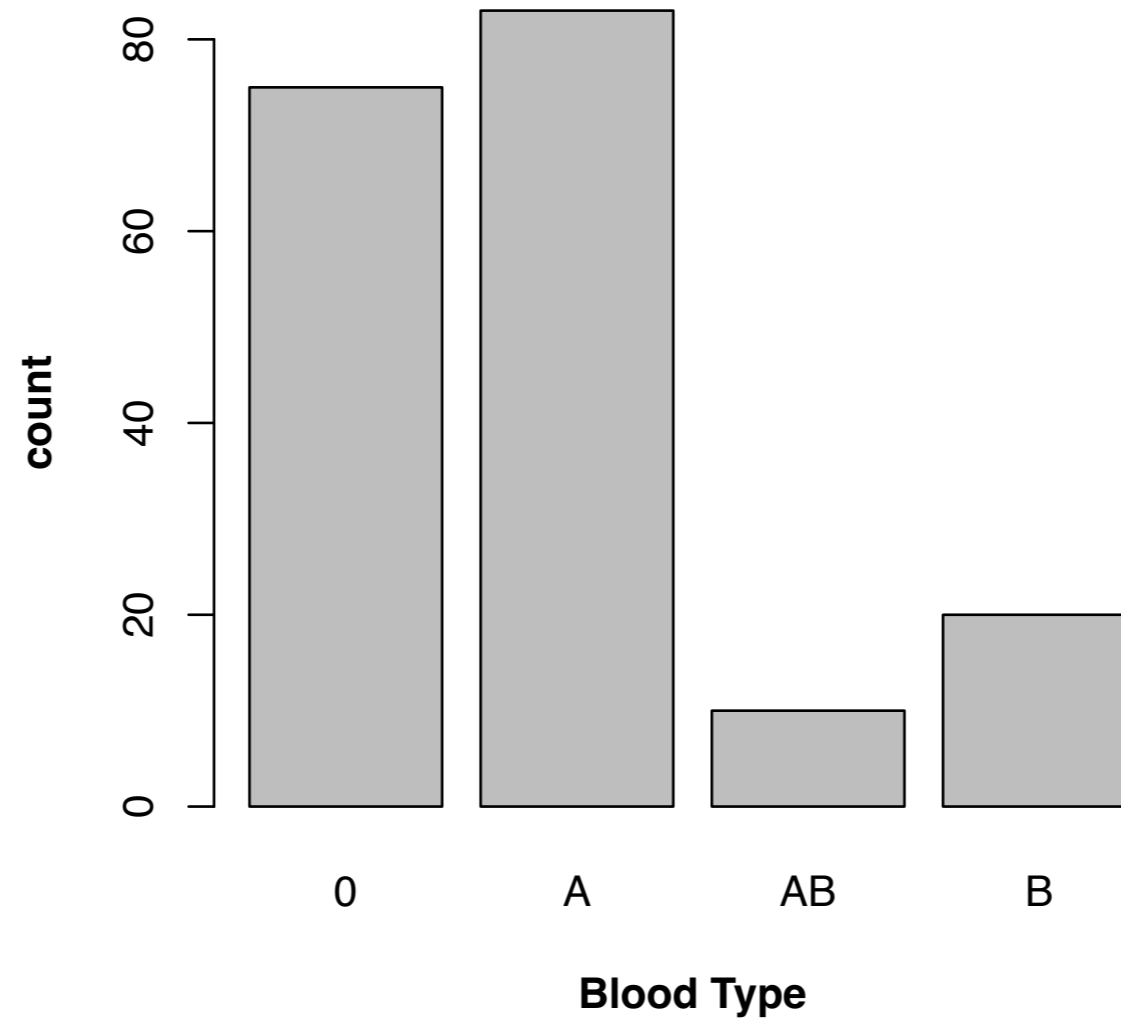
categorical variables

Categorical Variables - Table

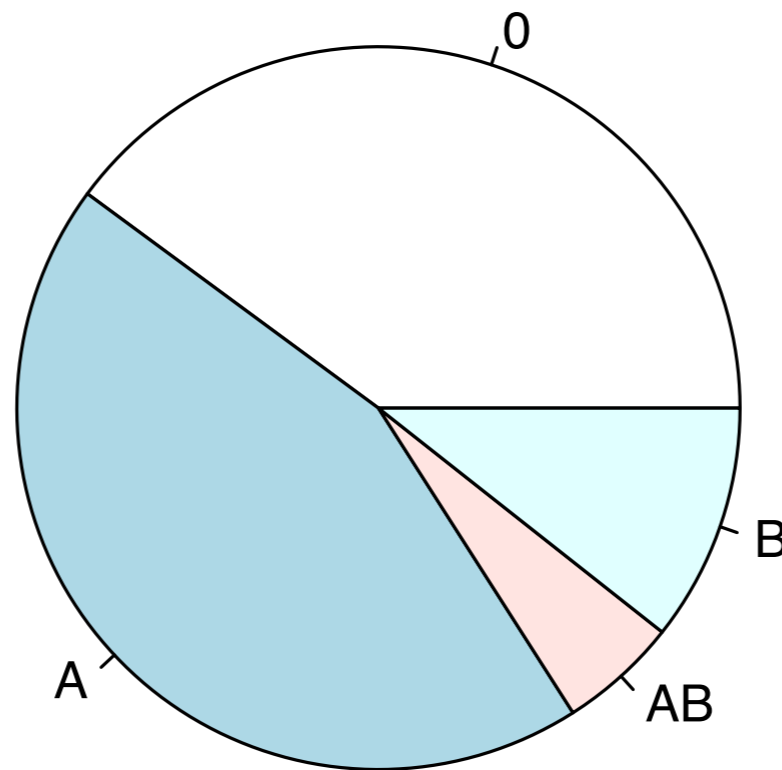
Value	A	B	AB	0	Σ
absolute frequency	75	83	10	20	188
relative frequency	40	44	5	11	100 %

n=188

Categorical Variables - Barplot



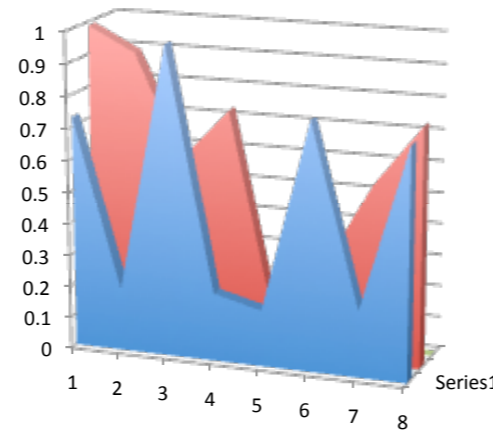
Categorical Variables - Piechart



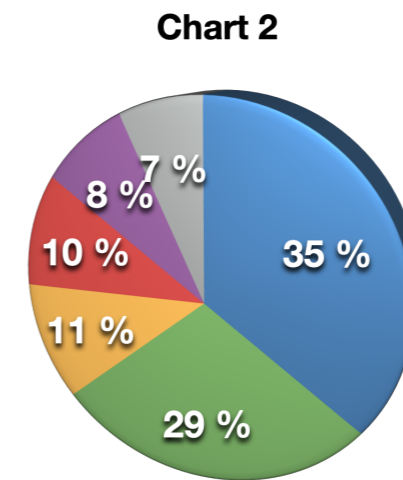
n=188

bad charts

- 3D displays

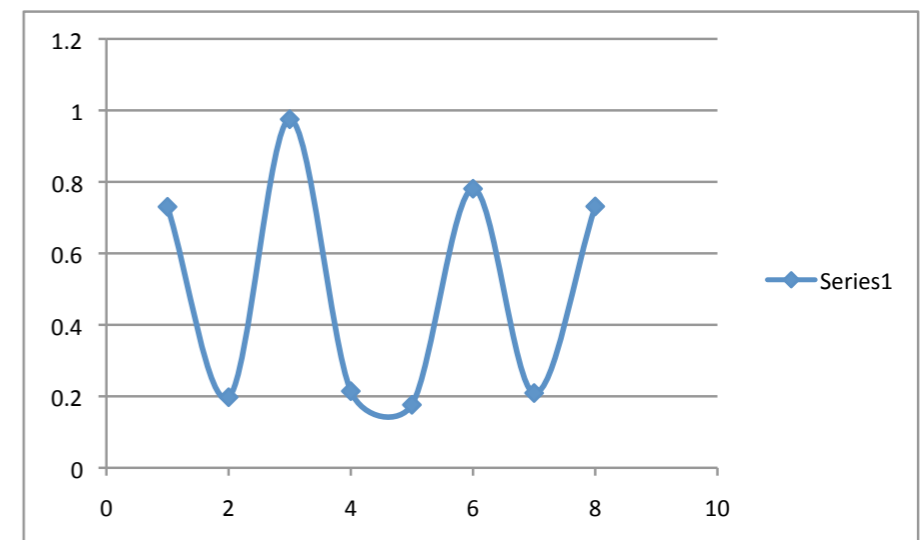


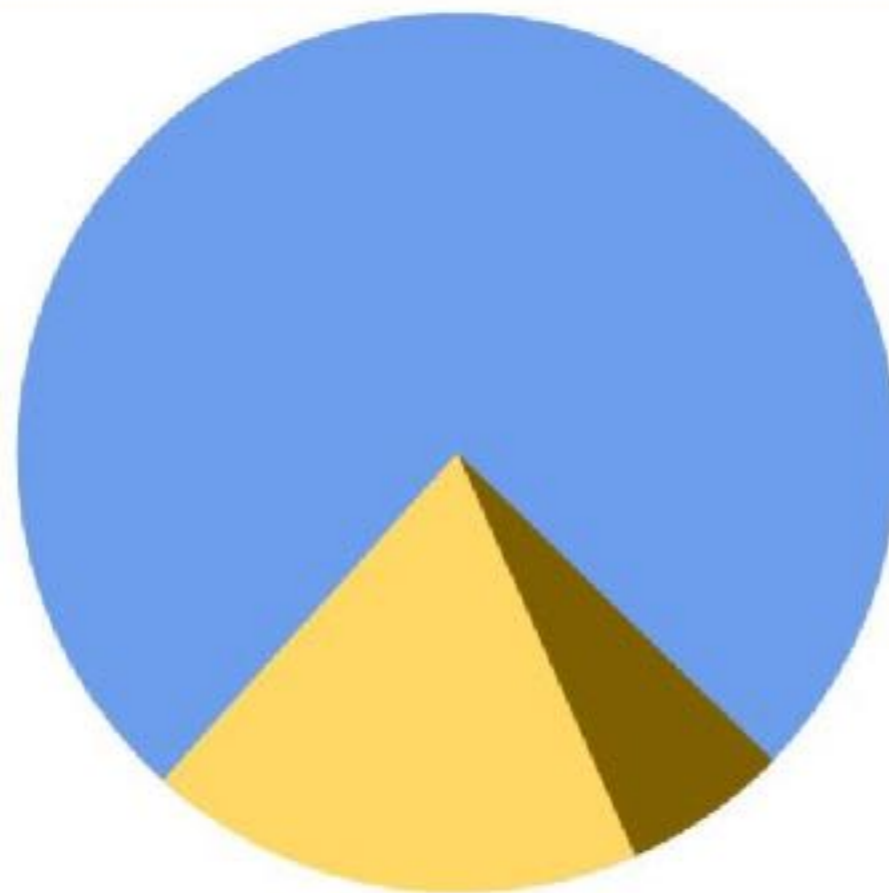
- pie charts
("the only place for a pice chart is a baker's convention")



● 2007 ● 2008 ● 2009 ● 2010 ● 2011 ● 2012

- smoothed curves in scatterplot, or any other lines in series, that are neither direct data point connectors nor based on an appropriate regression procedure





Sky



Sunny side of pyramid



Shady side of pyramid

Cross Tables - “Kontingenztafel”

Person	Medication	Response
A	verum	yes
B	pacebo	no

		Response	
		yes	no
Medication	verum	1	0
	placebo	0	1

effect

cause

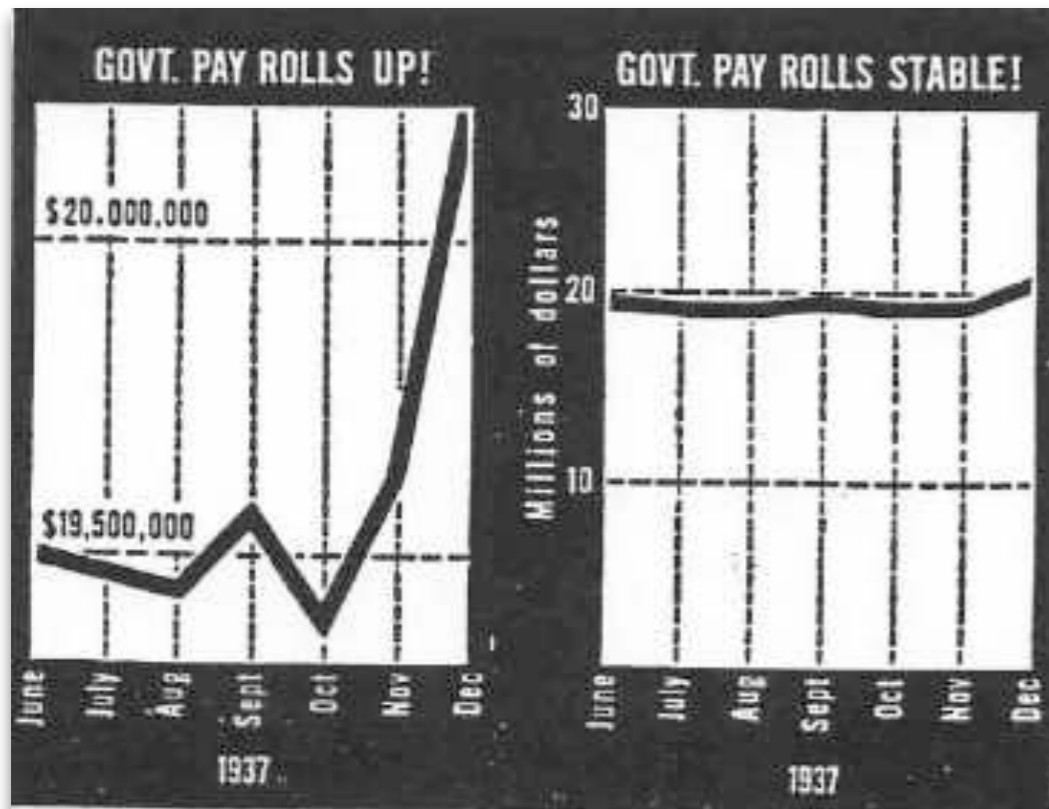
Cross Tables

n=80

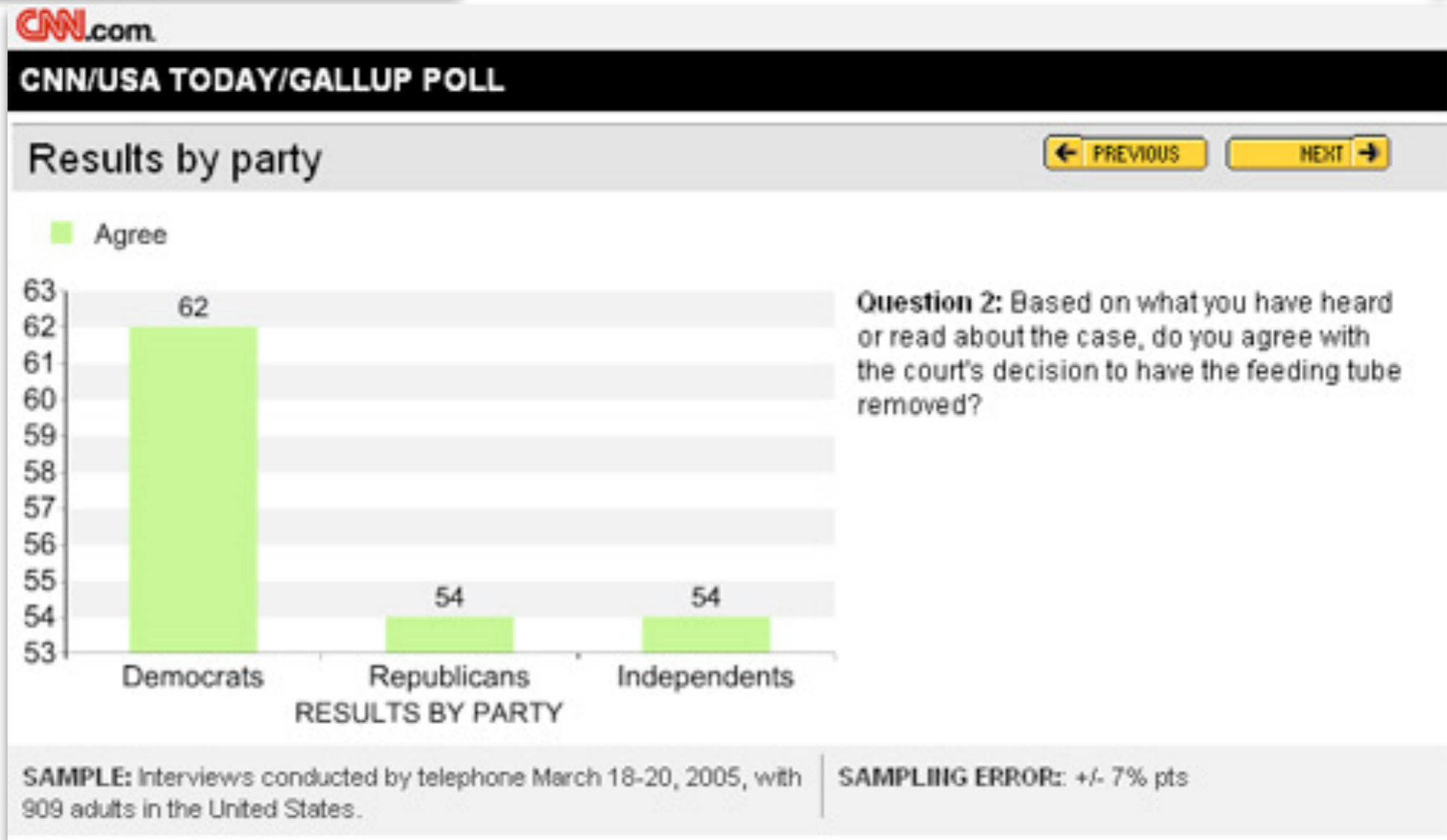
		Response		Total
		yes	no	
Medication	verum	20 50%,67%	20 50%,40%	40 50%
	placebo	10 25%,33%	30 75%,60%	40 50%
Total		30 37%	50 63%	80 100%

describing quantitative data

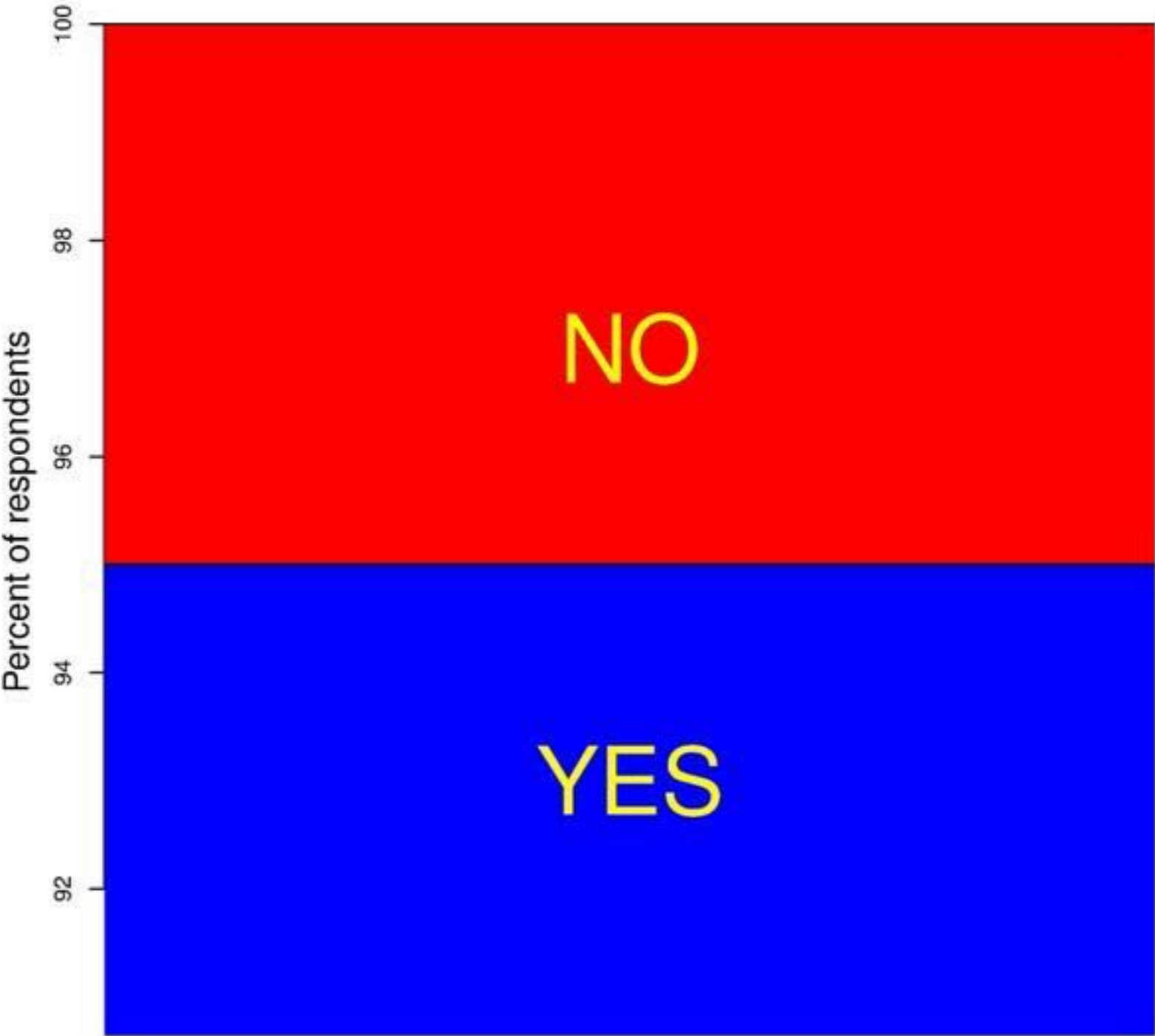
- Always report the sample size!
- numerical
median, Q1, Q3, min, max (5-point summary)
location and scatter (for symmetric distribution
mean, standard deviation)
- graphical
Histogram, Boxplot, Density Plots
- tables for categorical data
- verbal
*“mean BMI of Pima Indian females was 33.2 kg/m²
(n= 332, interquartile range = 28.2-37.2 kg/m²)”*

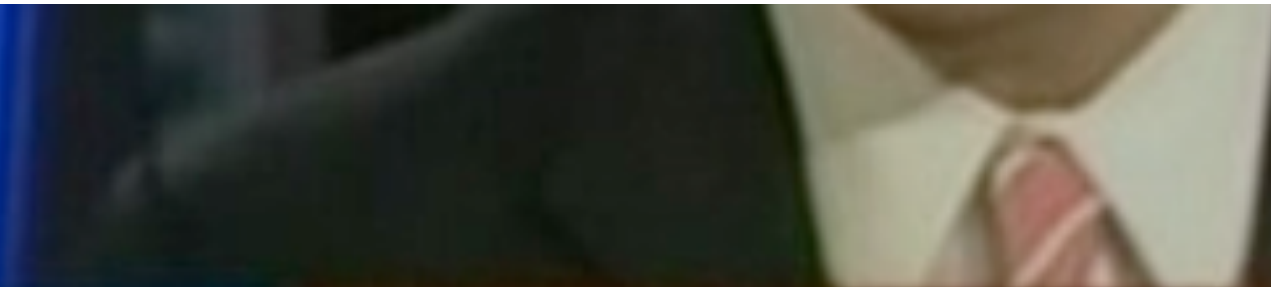


no sense is better suited for parallel processing than the visual sense. No sense has more built-in filters and processing steps. But it is the visual sense that can be fooled most easily.



Is truncating the Y-axis dishonest?





How did the President do?

A Excellent
84%

B Average
4%

C Poor
10%

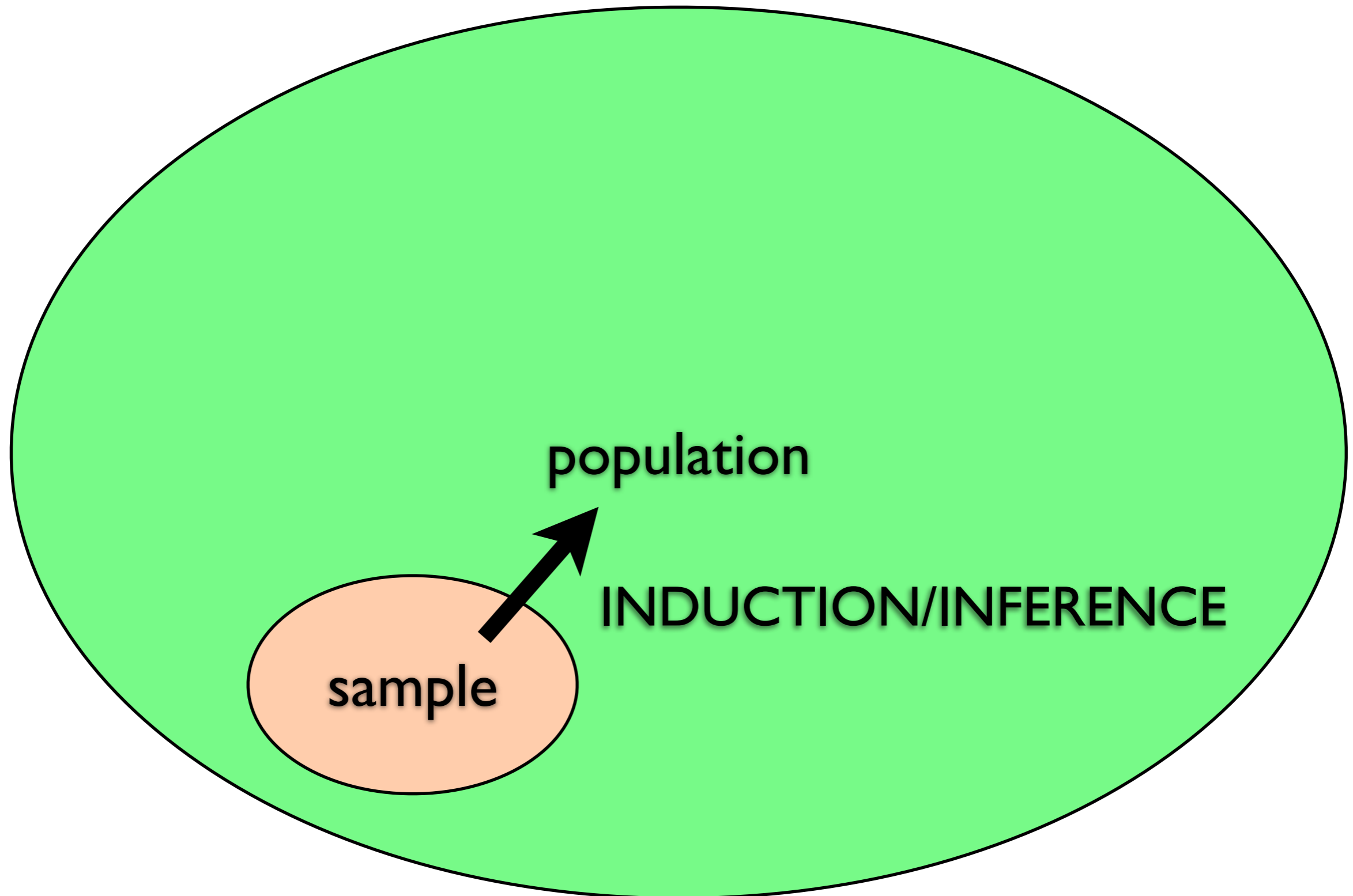
FOX
NEWS

STANDARD RATES APPLY. AVAILABLE ON MOST CELL CARRIERS.

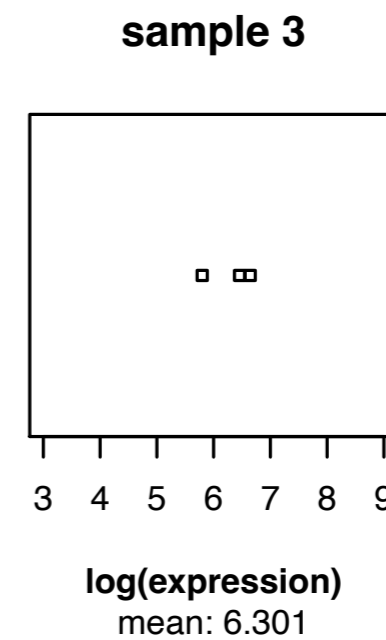
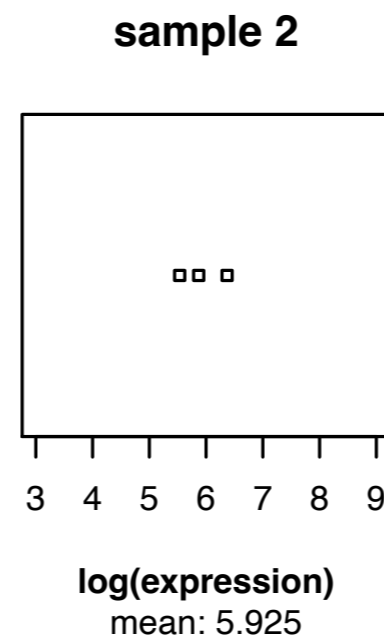
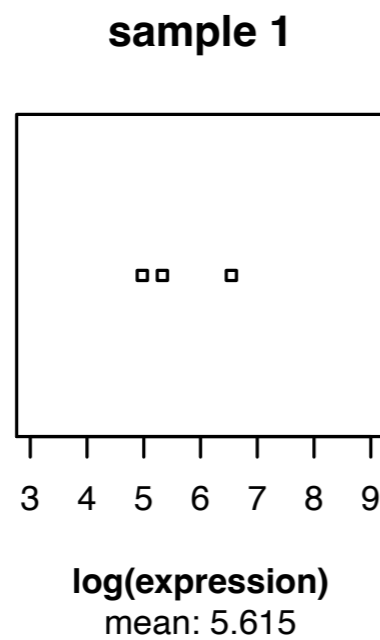
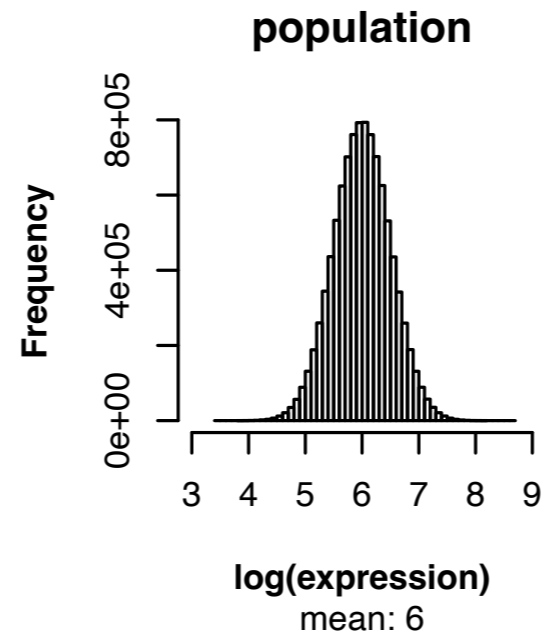
Inference

What really matters

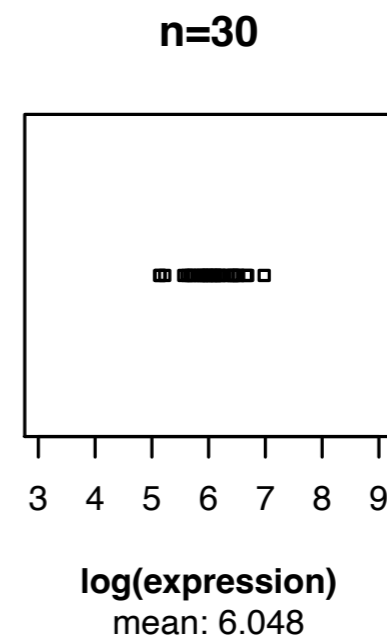
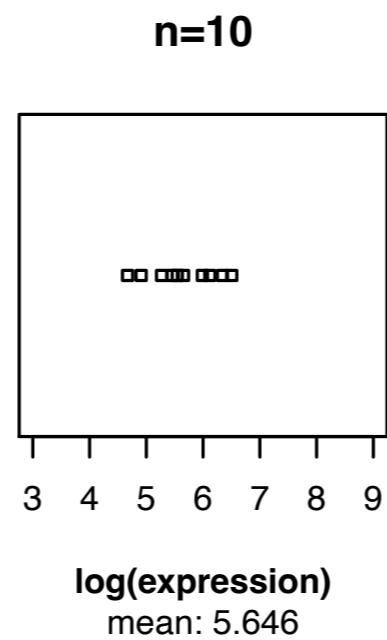
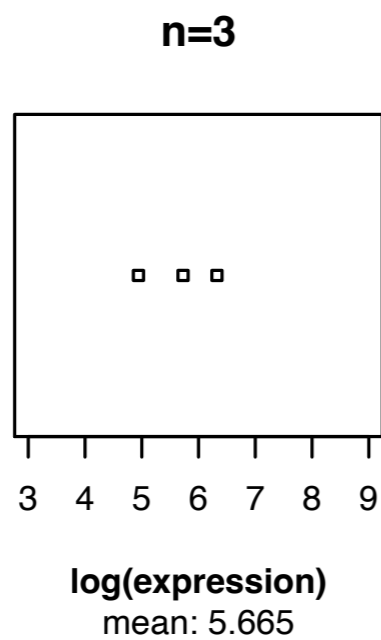
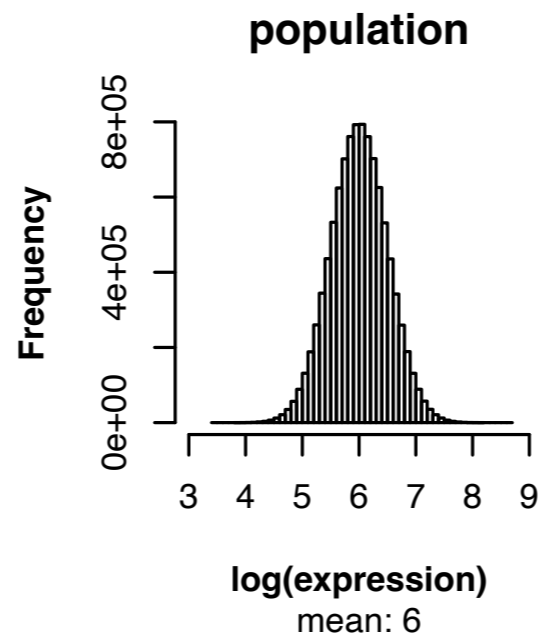
Sample - Population Relation



Sample - Population Relation

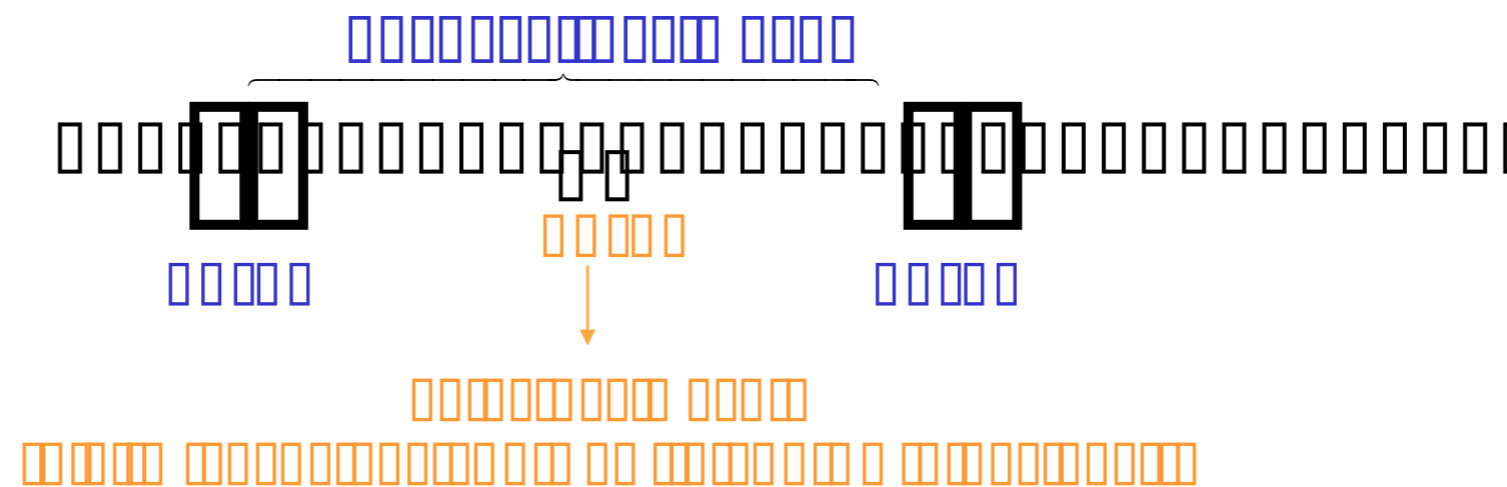


Sample - Population Relation



Confidence Intervals

- 95%-confidence interval: An estimated interval which contains the „true value“ of a quantity with a probability of 95%.



- $(1 - \alpha)$ -confidence interval: An estimated interval which contains the „true value“ of a quantity with a probability of $(1 - \alpha)$.

$1 - \alpha =$ confidence level, $\alpha =$ error probability

proportional data

You use a hemocytometer to determine the viability of cells stained with trypan blue. You count 94 unstained cells and 6 stained.

How can the data be represented?

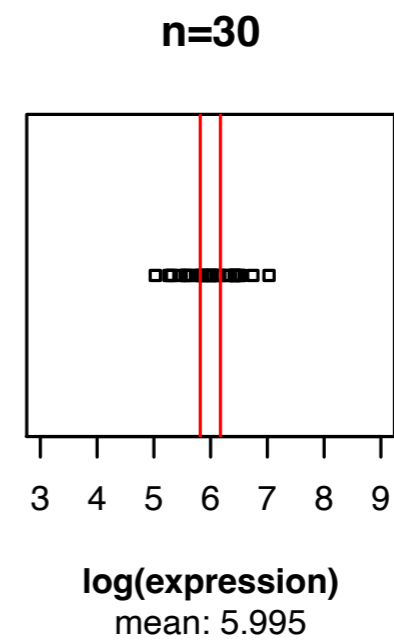
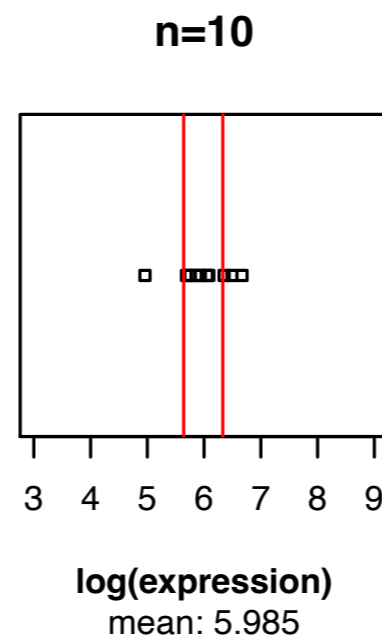
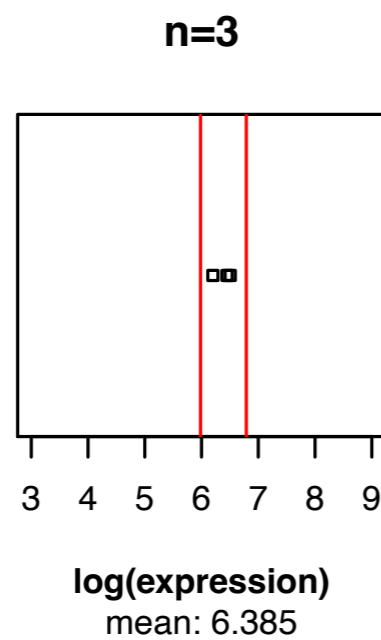
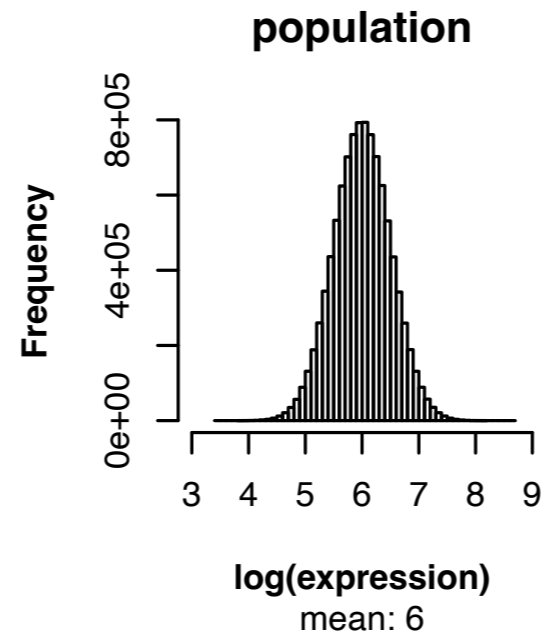
What is the 95% CI for the fraction of dead cells?

0.02 - 0.13 (binomial test, <http://statpages.org/confint.html>)

Which assumptions have to be made?

tube mixed well and the selection of sample was random

Confidence Intervals

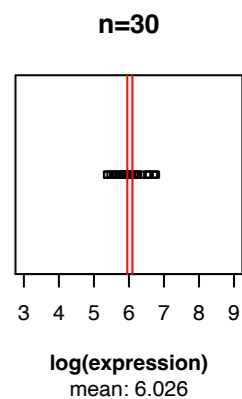
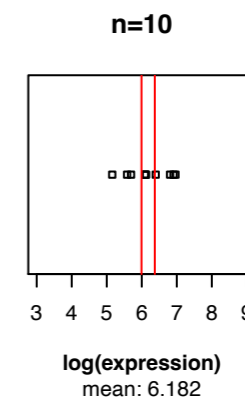
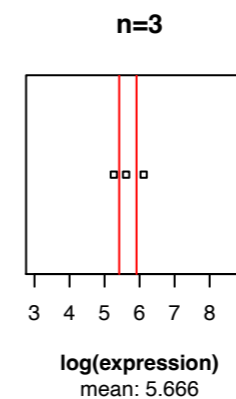
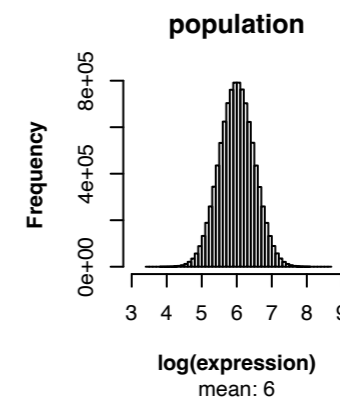


Standard Error of the Mean (SEM) - Standard Error

- The standard error of the mean (SEM) is the standard deviation of the sample mean estimate of a population mean.

$SEM = \text{standard deviation} / \text{square root}(n)$

- a small SEM indicates that the sample mean is likely to be quite close to the true population mean
- a large SEM indicates that the sample mean is likely to be far from the true population mean



Sample - Population

What allows us to conclude from the sample to the population?

The sample has to be **representative**

(Figures about drug abuse of students cannot be generalised to the whole population of Germany)

How is representativity achieved?

Large sample numbers

Random recruitment of samples from the population

Randomisation: Random allocation of the samples to the different experimental groups

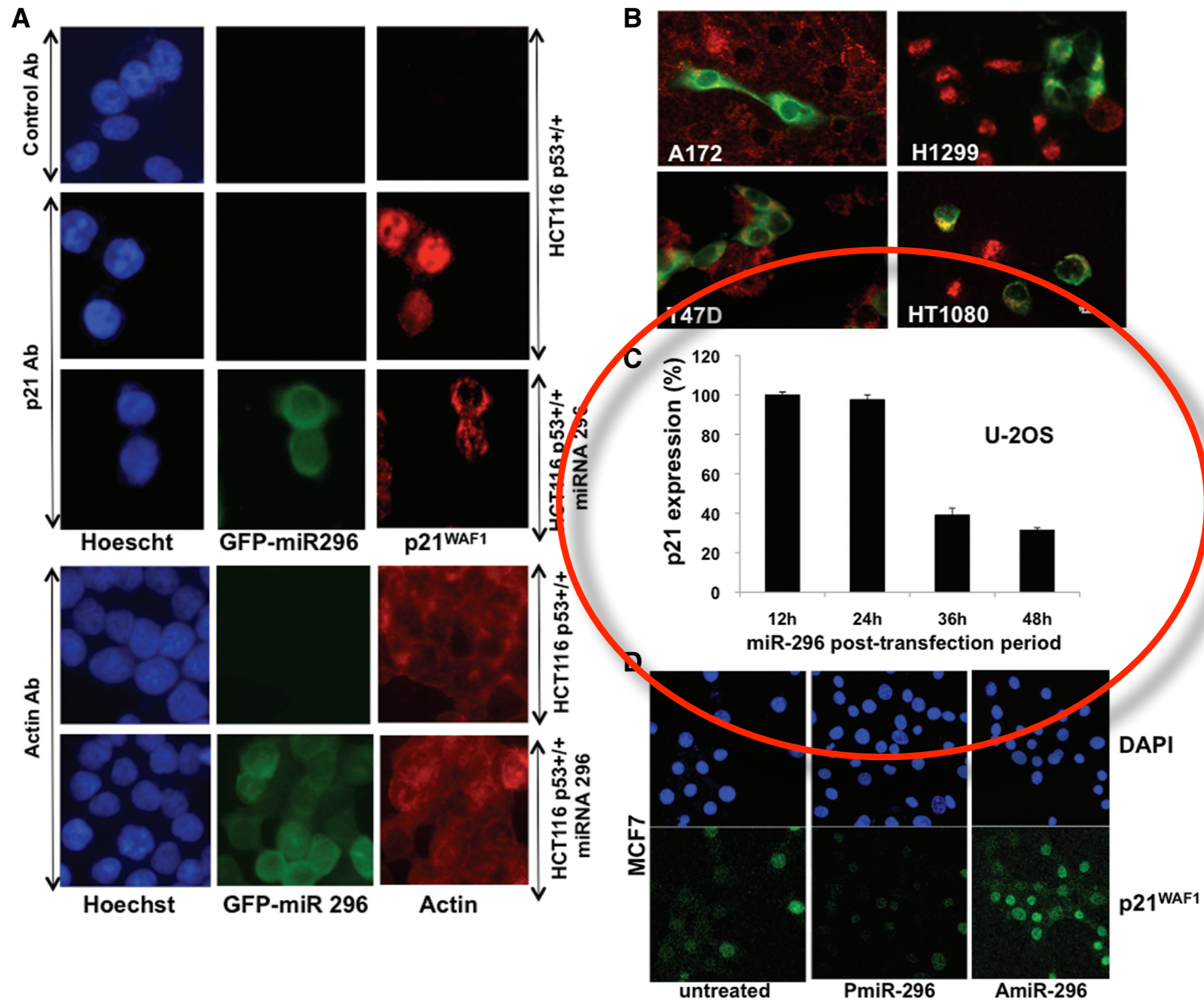


Figure 8. miR-296 downregulates p21 expression. (A) p21^{WAF1} expression (red) in control and pCXbG-miR-296 transfected cells. Secondary antibody (control Ab) and actin staining were used as negative controls. (B) A variety of cancer cells examined for p21^{WAF1} expression after transfection of miR-296 expression construct (green) showed lack of red staining in green cells demonstrating that miR-296 downregulates p21^{WAF1} expression. (C) Quantitative-PCR for p21^{WAF1} in U2-O S cells treated with PmiR-296 showed time-dependent decrease after treatment with PmiR-296. (D) Cells treated with PmiR-296 showing decrease and with AmiR-296 showing increase in p21^{WAF1} staining. The data demonstrate that the miR-296 regulates p21^{WAF1}.

Error Bars

Correspondence

Nature 428, 799 (22 April 2004) | doi:10.1038/428799c

Error message

David L.Vaux I

I. The Walter and Eliza Hall Institute, 1 G Royal Parade, Parkville, Victoria 3050, Australia

Sir

In the 19 February 2004 issue of Nature, there were **ten items** (one Brief Communication, one Article and eight Letters to Nature) containing figures **with error bars**, but **only three had figure legends describing what the error bars were**: in one case, 80% confidence intervals; in another, standard deviations; and in the third, standard error of the mean. The articles with incomplete legends represented both the biological and physical sciences, across many different disciplines, and clearly should not be considered isolated examples.

Error bars can be used by the reader to determine how much the data varied, allowing an estimation of whether the experiments gave reproducible results, whether the results were significantly different from the controls, and sometimes whether the data were obtained in an unbiased manner.

How did these omissions occur? If authors include error bars on their figures, why do they so often forget to state what they are in the legends? **How can reviewers be confident that the conclusions are correct if they are not told about the errors in the data?** Why don't reviewers request that descriptions of the error bars be included when they review the papers?

When properly described, error bars can be very revealing. In their analysis of the experiments and methods used by Jacques Benveniste to study homeopathy, John Maddox and colleagues illustrated how much information can be gained if one knows how to interpret errors correctly (Nature 334, 287–290; 198810.1038/334287a0).

By not ensuring that all papers that have error bars describe what they are, Nature publishes material that cannot be properly assessed by its readers.

Nature is fortunate in having such attentive readers. Our editors and reviewers expect error bars to be properly defined, and we shall be more vigilant in ensuring best practice in future — Editor, Nature.

Errors Bars - which and when

- show **SD** when you are interested in showing the *scatter*
- show the **SEM** (or confidence interval) when you want to know *how well you know the population mean*
- some people like to display SEM for another reason: SEMs are smallest measure of error and thus look nicest (SEM = SD/SQRT(n)) **always report n!**
- The scatter (however expressed) means different things in different contexts. Is the author showing the variability among replicates in a single experiment? Variability among experiments with genetically identical animals? Variability among cloned cells, or within patients? etc. etc.

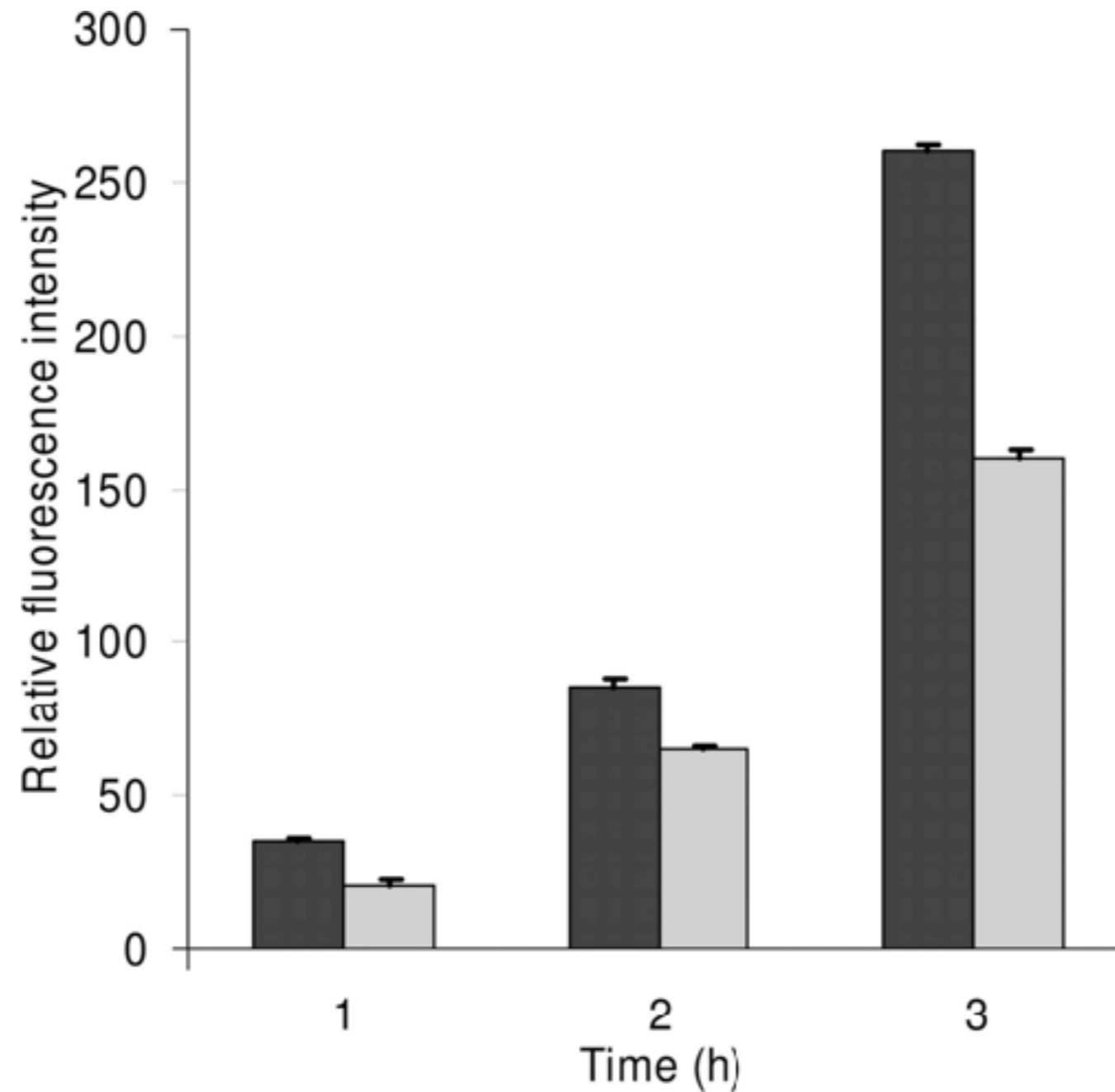


Figure 3. Enzyme activity for MEFs showing mean + SD from duplicate samples from one of three representative experiments. Values for wild-type vs. -/- MEFs were significant for enzyme activity at the 3-h timepoint ($P < 0.0005$).

A Journal's "Rules"

- the value of n (i.e., the sample size, or the number of independently performed experiments) must be stated in the figure legend.
- error bars and statistics should only be shown for independently repeated experiments, and *never* for technical replicates. If a "representative" experiment is shown, it should not have error bars or P values, because in such an experiment, $n = 1$
- because experimental biologists are usually trying to compare experimental results with controls, it is usually appropriate to show inferential error bars, such as SE or CI, rather than SD. However, if n is very small (for example $n = 3$), rather than showing error bars and statistics, it is better to simply plot the individual data points.

the link between error bars and significance

- The link between error bars and statistical significance is weaker than many wish to believe.
- But: if two SEM error bars overlap you can conclude that the difference is not statistically significant ($p > 0.05$), but that the converse is not true.
- Some graphs and tables show the mean with the standard deviation (SD) rather than the SEM. The SD quantifies variability, but does not account for sample size. To assess statistical significance, you must take into account sample size as well as variability.
Therefore, observing whether SD error bars overlap or not tells you nothing about whether the difference is, or is not, statistically significant.

case study

measurements

An enzyme level is measured in cultured cells. The experiment is repeated on 3 days. Each day triplicate measurements (technical replications) are performed.

Summarize the data and justify your procedure

	replicate 1	replicate 2	replicate 3
Monday	234	220	229
Tuesday	269	967	275
Wednesday	254	249	246

units/(min*mg)

measurements

	replicate 1	replicate 2	replicate 3	Mean
Monday	234	220	229	227,67
Tuesday	269	967	275	272
Wednesday	254	249	246	249,67
Grand Mean				249,78

units/(min*mg)

“The experiment was performed three times in triplicate. After removing one extreme outlier, the mean for each experiment was calculated. The grand mean is 249.8. The 95% CI ranges from 194.7 to 304.9. (n=3)”

Descriptive Stats - Best of

- Publish all raw data
- Summarise sensibly
- Report N
- Inference matters
- Don't trust your eyes

Imputation

Why Missing Values?

- MCAR: missing completely at random
- MAR: missing at random
missing-ness can be predicted
- NMAR: not missing at random
correlation with unobservable
characteristic

Strategies to deal with missing values

- List-wise deletion (>5% dropout)
- Pairwise deletion
- Mean/Median substitution
- Multiple imputation

Multiple imputation

- Impute
- Repeat 3-5 times
- Perform desired analysis on each repetition
- Average parameter estimates to obtain single point estimate
- Calculate SE based on variation across datasets

stats twitter accounts to follow

- @d_spiegel
- @statsepi
- @MaartenvSmeden
- @lakens

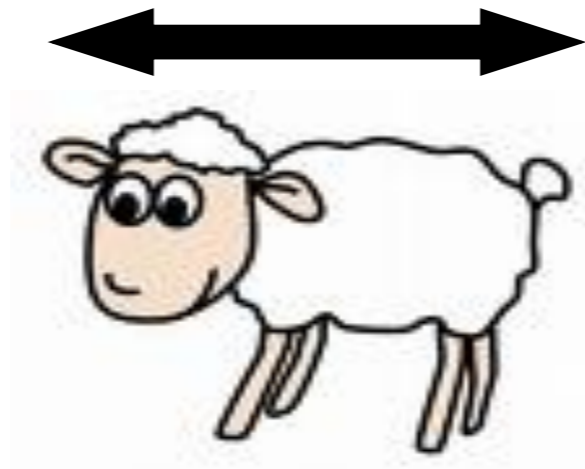
- @VPrasadMDMPH
- @ProfDFrancis

Test Theory

non-sheep detector

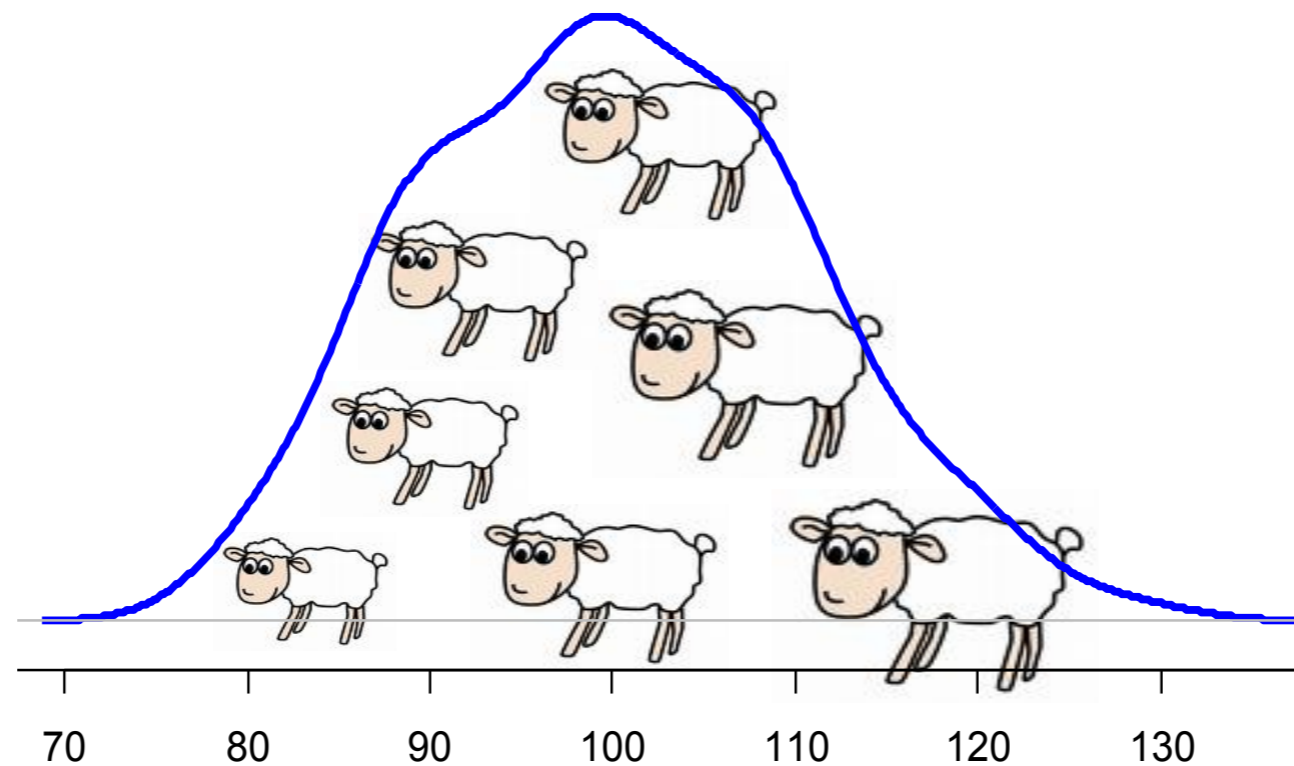
Training:

Measure the length of all sheep that cross your way



non-sheep detector

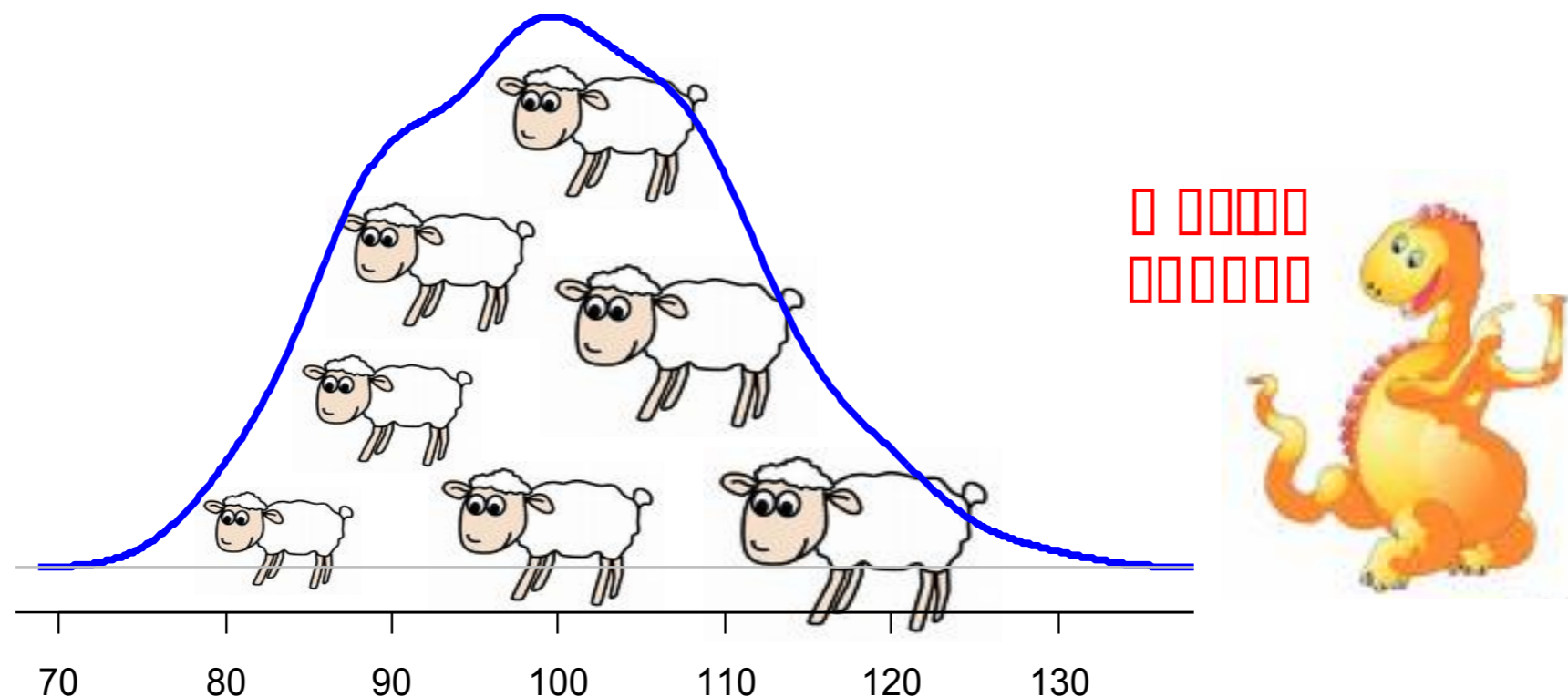
Determine the distribution of the quantity of interest
(length of sheep).



non-sheep detector

Test phase:

For any unknown animal, test the hypothesis that it is a sheep. Measure its length and compare it to the learned length distribution of the sheep. If its length is „out of bounds“, the animal will be called a non-sheep (rejection of the hypothesis). Otherwise, we cannot say much (non-rejection).



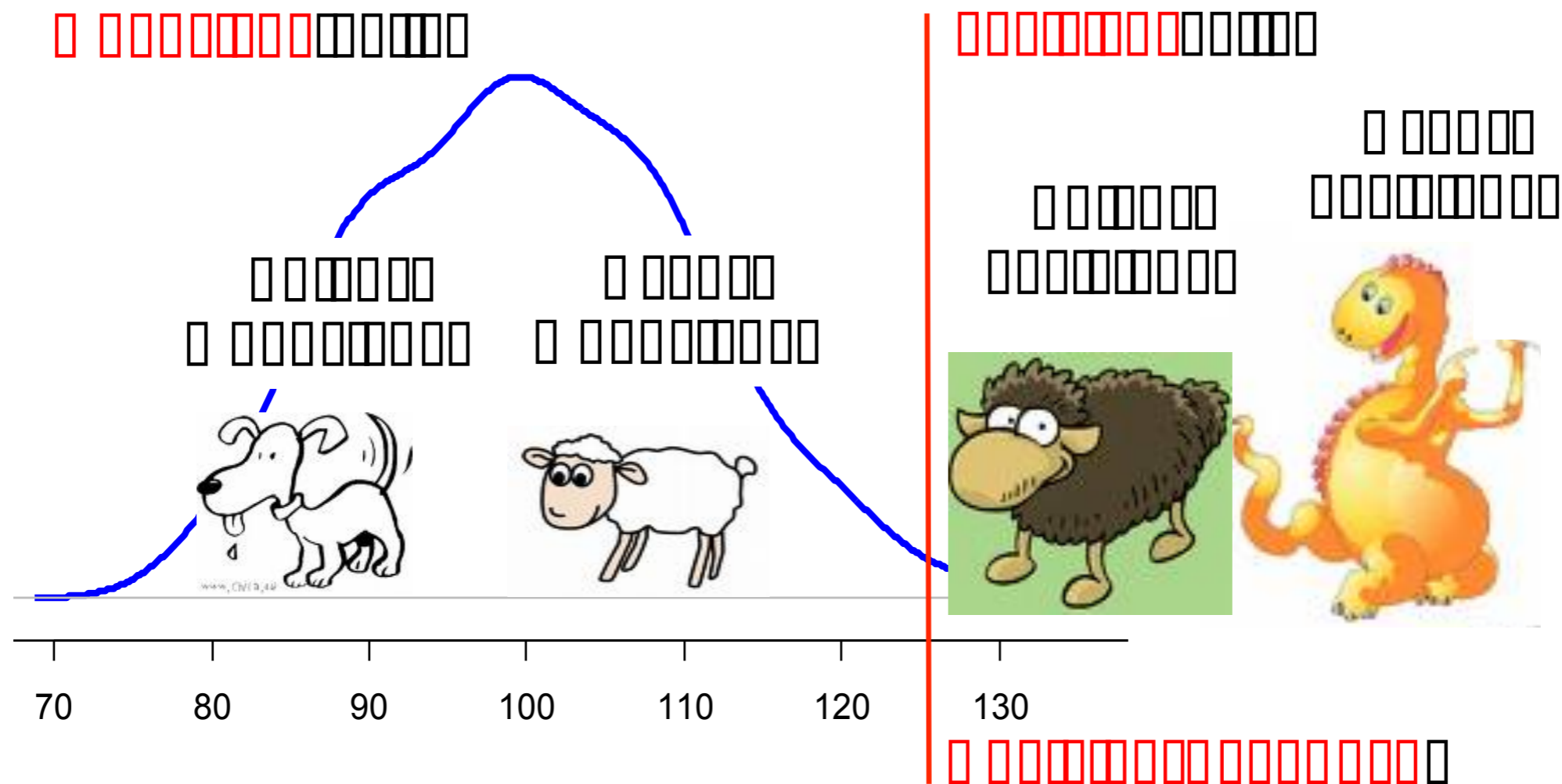
non-sheep detector

Advantage of the method:

One does not need to know much about sheep.

Disadvantage:

It produces errors...



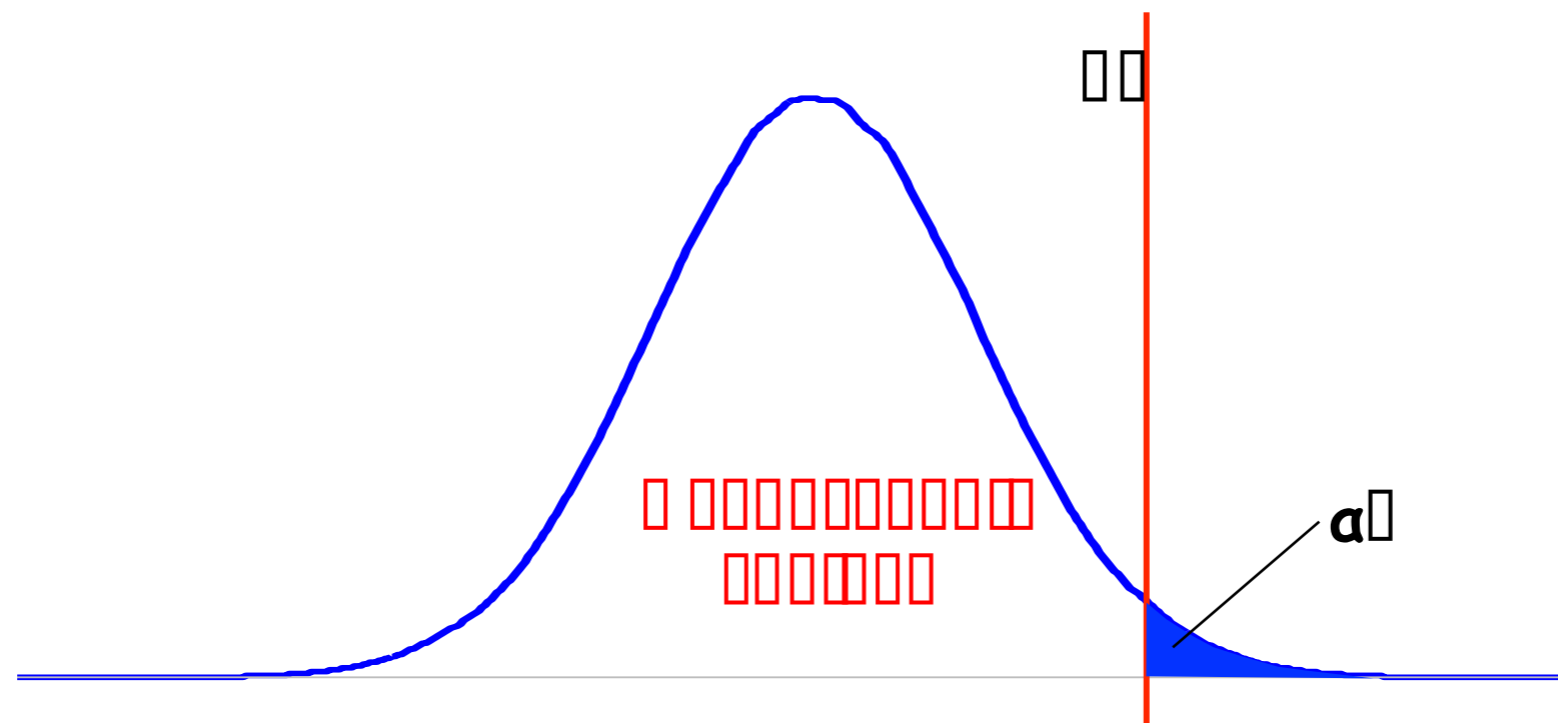
Statistic Hypothesis Testing

- State a **null hypothesis** H_0
 (“nothing happens, there is no difference...”)
- Choose an appropriate **test statistic** (the data-derived quantity that finally leads to the decision)
 This implicitly determines the null distribution (the distribution of the test statistic under the null hypothesis).



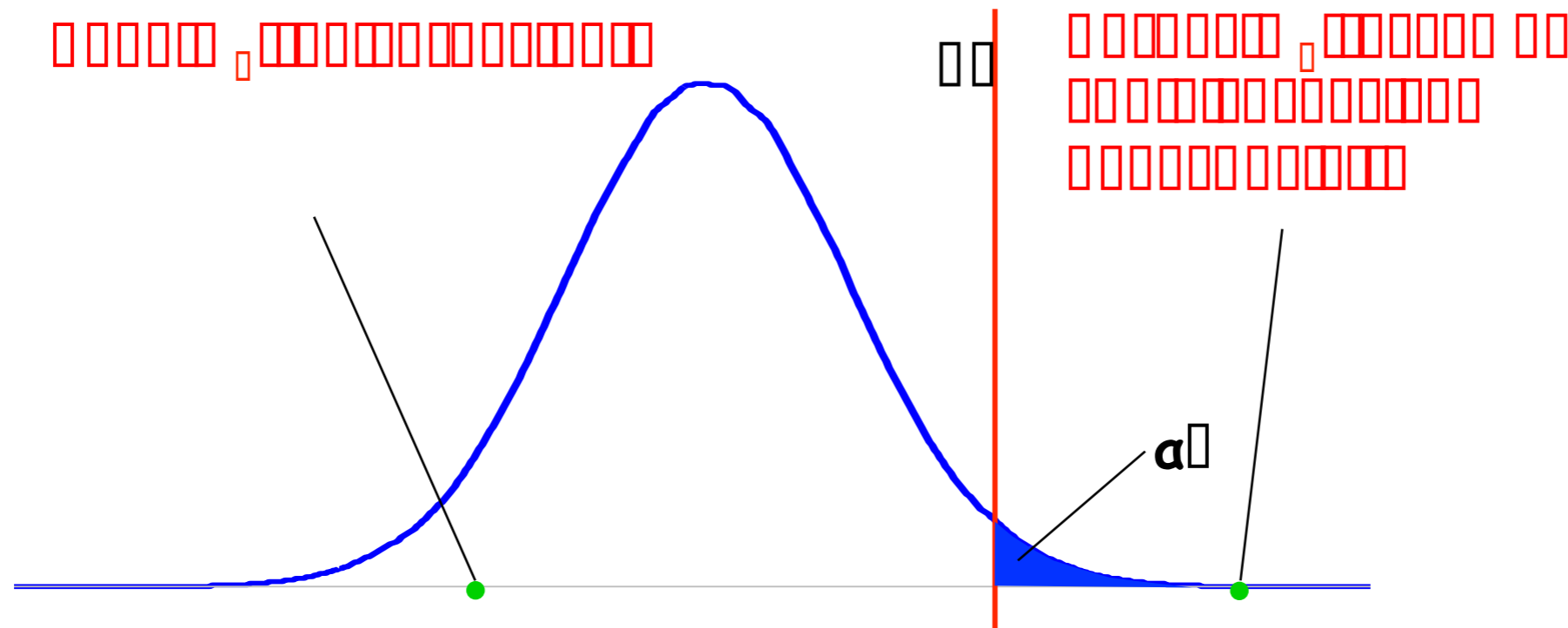
Statistic Hypothesis Testing

- State an **alternative hypothesis** (e.g. “the test statistic is higher than expected under the null hypothesis”)
- Determine a **decision boundary**. This is equivalent to the choice of a **significance level α** , i.e. the fraction of false positive calls you are willing to accept.

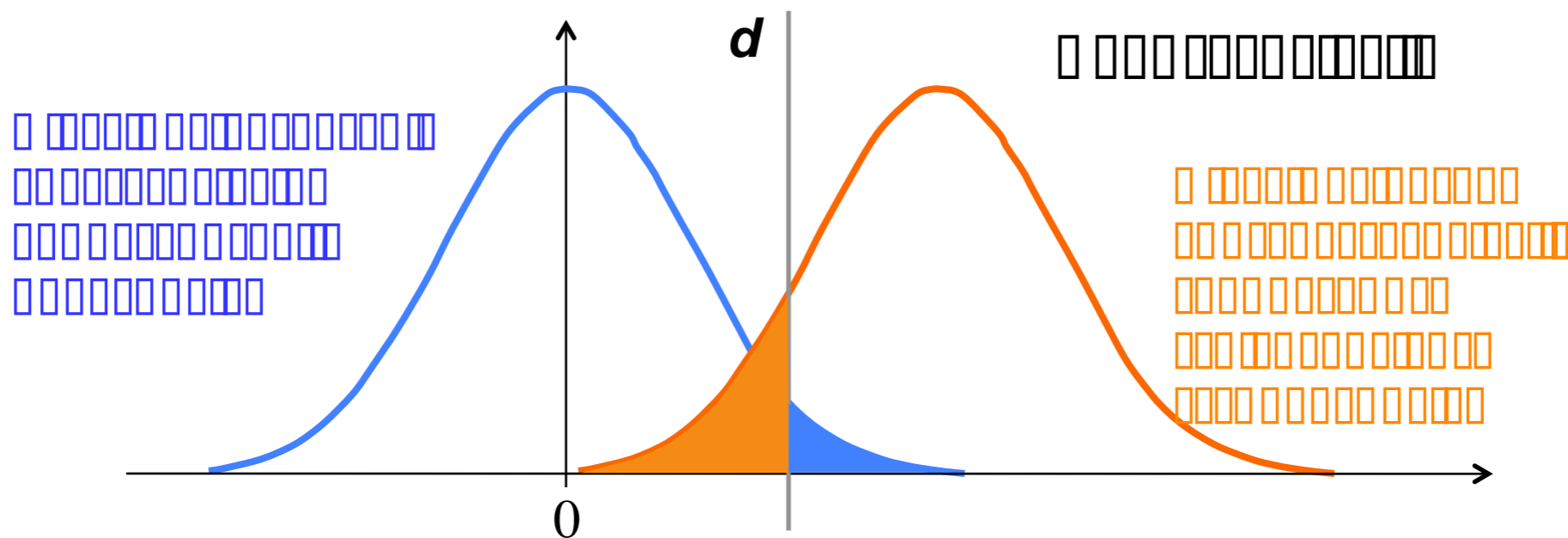


Statistic Hypothesis Testing

- Calculate the actual **value** of the test statistic in the sample, and make your decision according to the pre-specified(!) decision boundary.

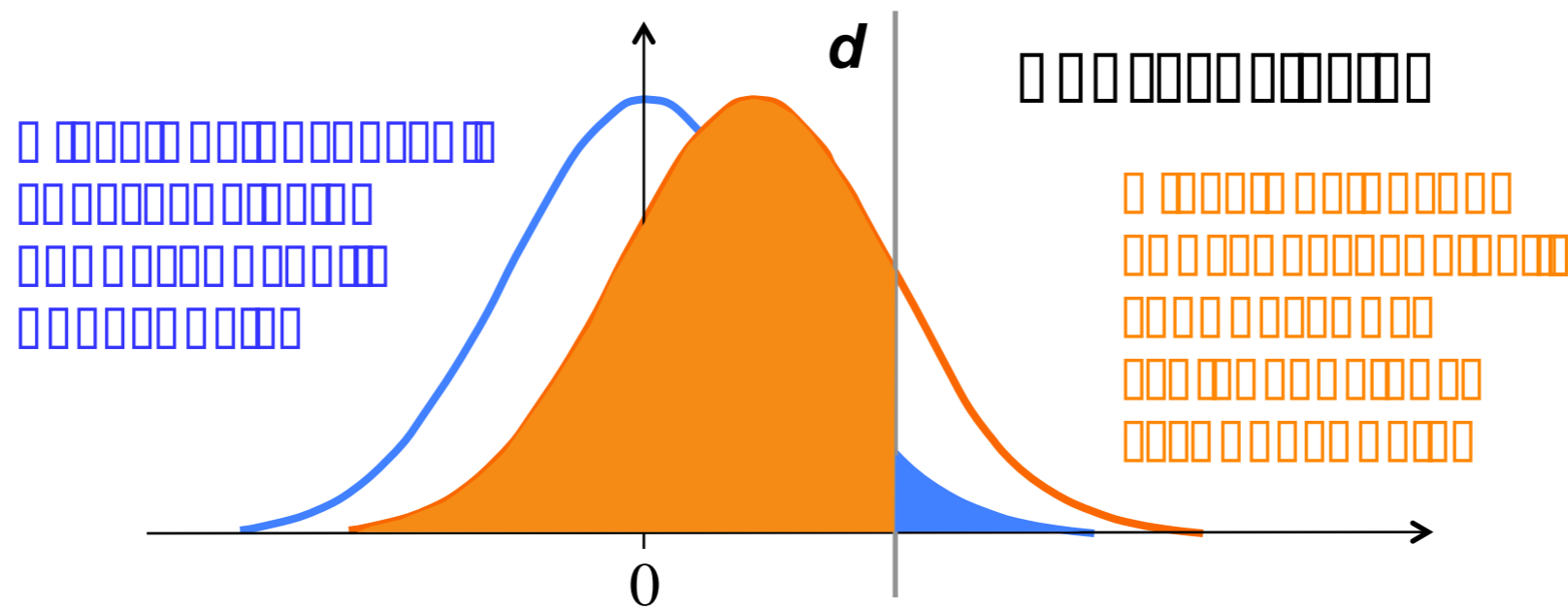


Good/Bad Test Statistics



	Accept null hypothesis	Reject null hypothesis
null hypothesis is TRUE	correct decision	Type I Error "False Positive"
alternative hypothesis is TRUE	Type II Error "False Negative"	correct decision

Good/Bad Test Statistics



	Accept null hypothesis	Reject null hypothesis
null hypothesis is TRUE	correct decision	Type I Error "False Positive"
alternative hypothesis is 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 (i.e. the probability of not committing a Type II error).
- As the power increases, the chances of a Type II error occurring decrease. The probability of a Type II error occurring is referred to as the false negative rate (β). Therefore power is equal to $1 - \beta$, which is also known as the sensitivity.

IMPORTANT!!

- Statistical Power = $1 - \beta$
- It is wrong to assume that type I error (false positives) rates are independent of the power.

In fact, it has been shown that many (most) significant results published are false positives also thanks to low statistical power of the test applied

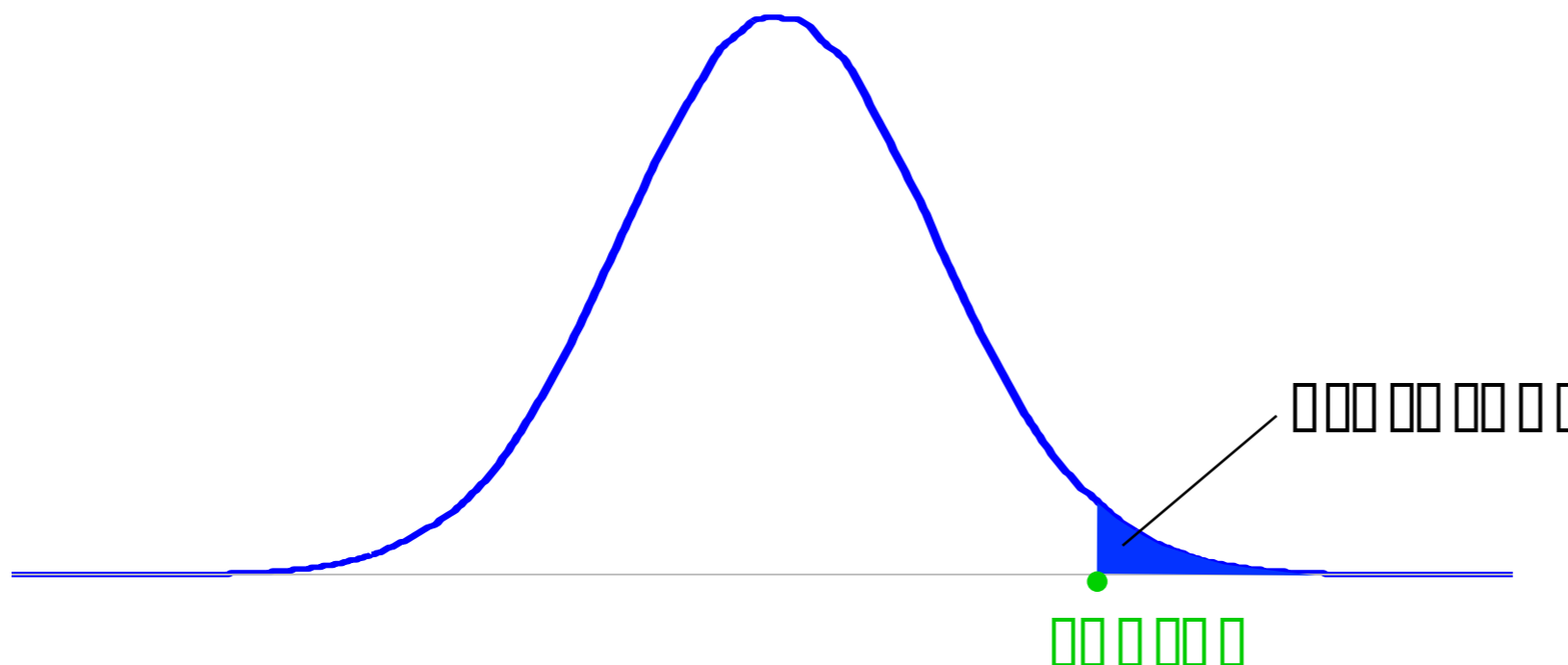
Power analysis

- Goal is to allow you to decide, while in the process of designing an experiment,
 - (a) **how large a sample** is needed to enable statistical judgments that are accurate and reliable and
 - (b) **how likely** your statistical test will be to **detect** effects of a given size in a particular situation.
- Performing power analysis and sample size estimation is an important aspect of experimental design, because without these calculations, sample size may be too high or too low. If sample size is **too low**, the experiment will lack the precision to provide reliable answers to the questions it is investigating. If sample size is **too large**, time and resources will be wasted, often for minimal gain.

The p-value

Given a **test statistic** and its actual **value t** in a sample, a p-value can be calculated:

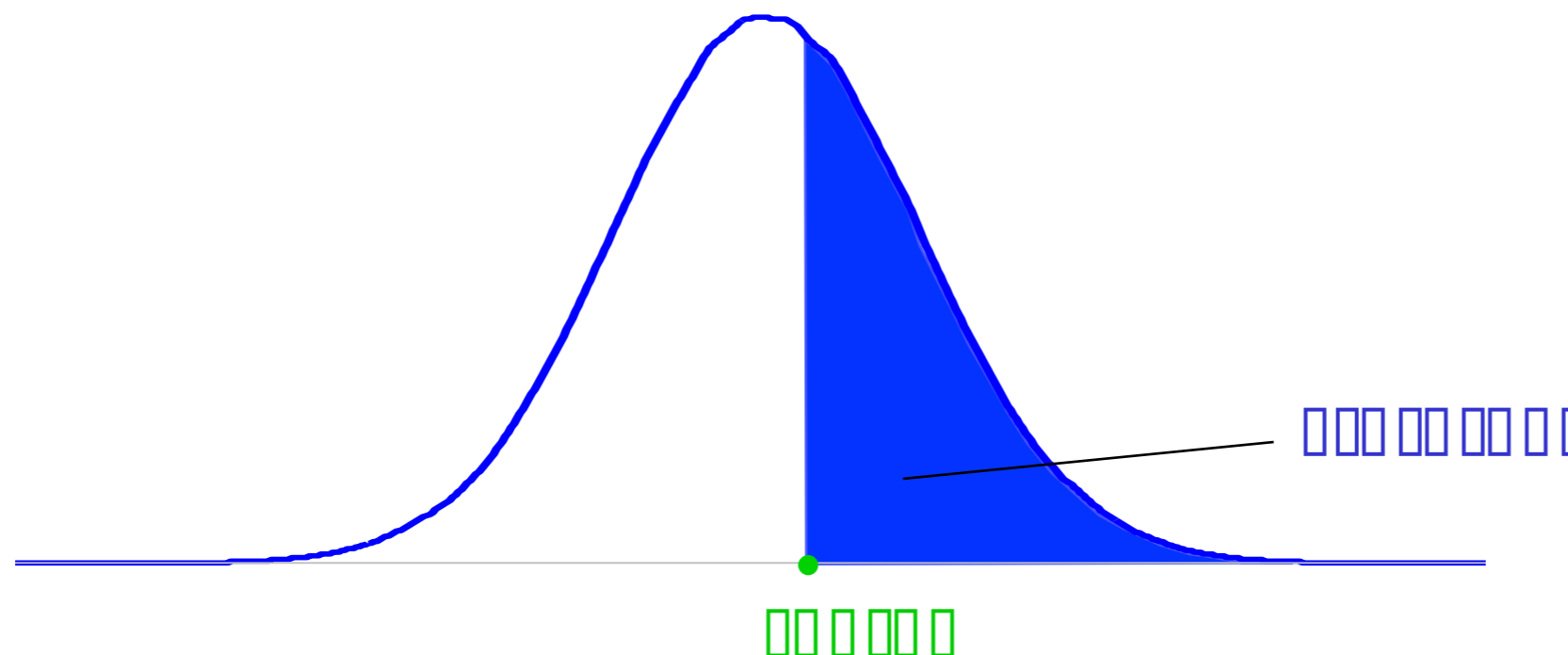
Each **test value t** maps to a p-value, the latter is the probability of observing a value of the test statistic which is at **least as extreme** as the actual value t (under the assumption of the null hypothesis).



The p-value

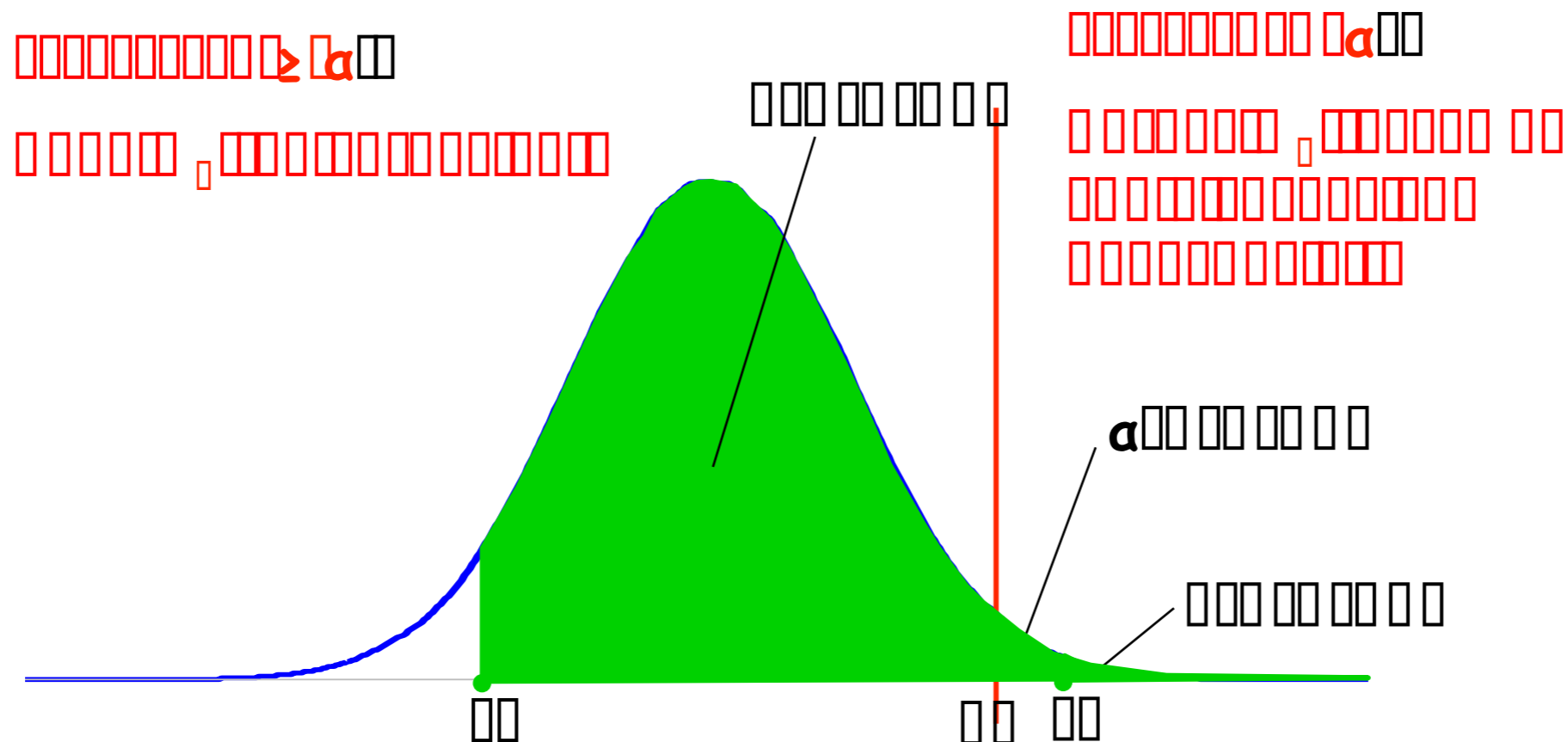
Given a **test statistic** and its actual **value t** in a sample, a p-value can be calculated:

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



Test decisions according to p-value

Decision boundary d	\longleftrightarrow	significance level α
Observed test statistic t	\longleftrightarrow	p-value
t more extreme than d	\longleftrightarrow	p smaller than α



$p > a$ does
not!

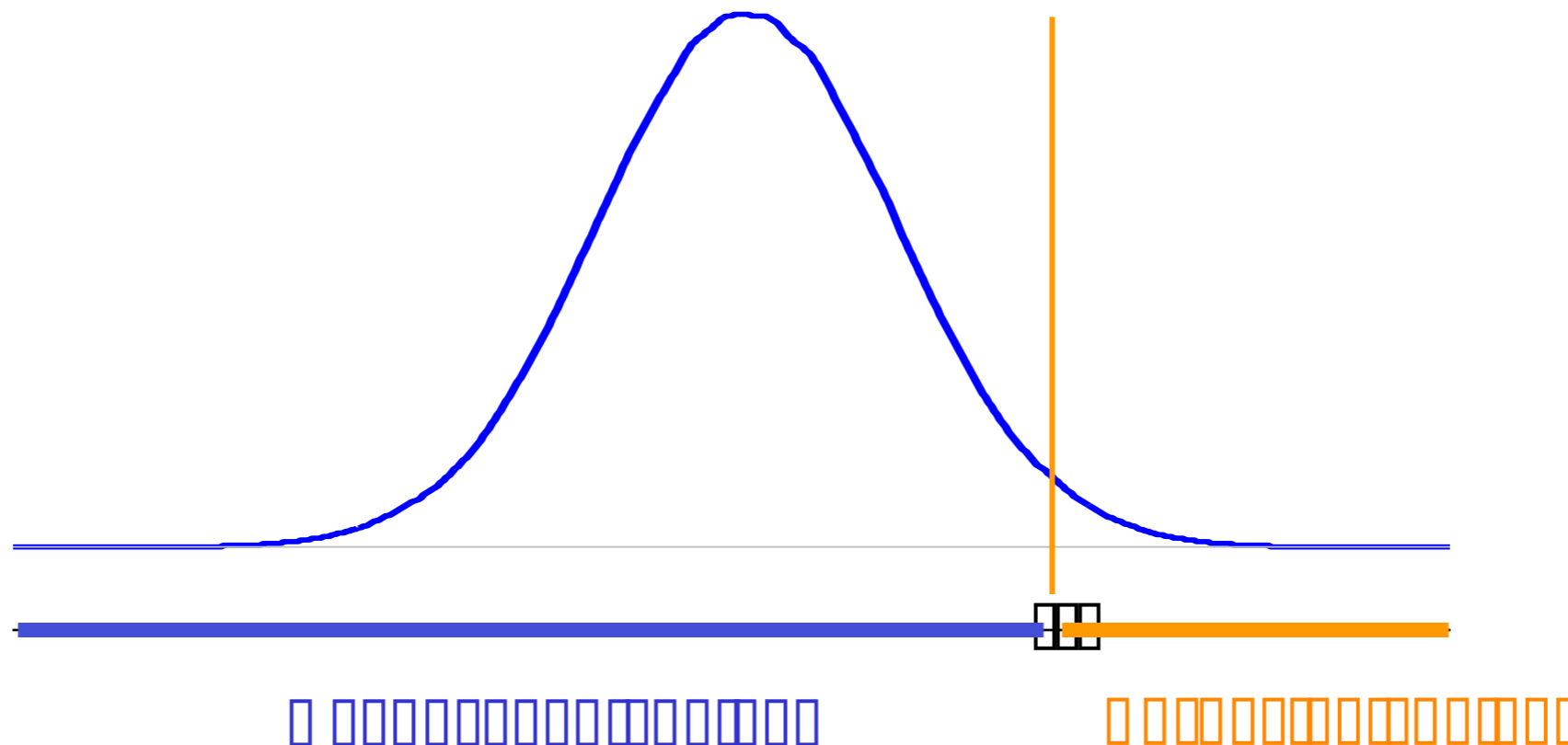
prove equality

one- and two-sided hypotheses

one-sided alternative

H_0 : The value of a quantity of interest in group A is not higher than in group B.

H_1 : The value of a quantity of interest in group A is higher than in group B



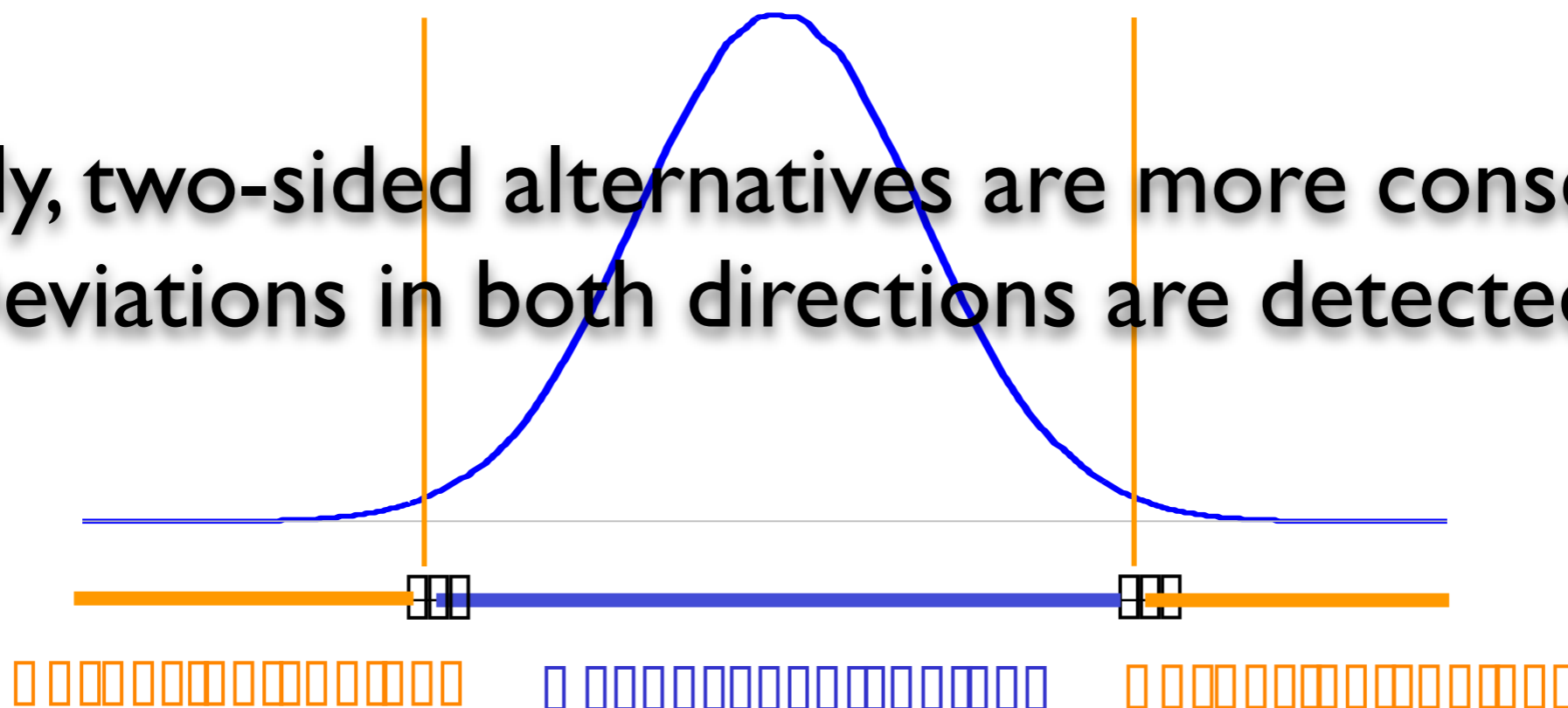
one- and two-sided hypotheses

two-sided alternative

H₀: The quantity of interest has the *same value* in group A and group B

H₁: The quantity of interest is *different* in group A and group B

Generally, two-sided alternatives are more conservative:
Deviations in both directions are detected.



Neuer Impfstoff aktiviert das Immunsystem

Erfolg gegen Krebs

USA: Überlebensrate der Patienten erhöht

HEIDELBERG – Amerikanischen Wissenschaftlern ist es erstmals gelungen, das Immunsystem von Krebspatienten mit einem Impfstoff direkt zu aktivieren und so die Überlebensrate der Erkrankten zu erhöhen.

Auf einem Symposium der Deutschen Krebsgesellschaft stellten die US-Mediziner Mike Hanna und William Cassel die aufsehenerregenden Ergebnisse ihrer Forschung vor. An den Versuchen nahmen insgesamt **62 Patienten teil**, die alle an Dickdarmkrebs erkrankt waren und bereits Tochtergeschwülste ausgebildet hatten. **32** der von Professor Hanna operierten Personen erhielten einen Monat nach dem Eingriff **Impfstoff gespritzt**, der aus Zellen ihres Tumors und aus einem Stoff, der das Immunsystem stimuliert, besteht. Die Tumorzellen waren vorher mit Strahlen behandelt worden, um sie unschädlich zu machen.

Zum Erstaunen der Ärzte zeigten fast alle Patienten eine Reaktion des Immunsystems. **Nach vier Jahren lebten von den 32 Versuchspersonen noch 94 Prozent.** Bei einer Kontrollgruppe, die nicht geimpft worden war, waren **noch 77 Prozent am Leben.** Durch die Injektion traten auch weniger Zweitumore auf als bei der Kontrollgruppe.

Colon carcinoma test

Neuer Impfstoff aktiviert das Immunsystem

Erfolg gegen Krebs

USA: Überlebensrate der Patienten erhöht

HEIDELBERG – Amerikanischen Wissenschaftlern ist es erstmals gelungen, das Immunsystem von Krebspatienten mit einem Impfstoff direkt zu aktivieren und so die Überlebensrate der Erkrankten zu erhöhen.

Auf einem Symposium der Deutschen Krebsgesellschaft stellten die US-Mediziner Mike Hanna und William Cassel die aufsehenerregenden Ergebnisse ihrer Forschung vor. An den Versuchen nahmen insgesamt **62 Patienten teil**, die alle an Dickdarmkrebs erkrankt waren und bereits Tochtergeschwülste ausgebildet hatten. **32** der von Professor Hanna operierten Personen erhielten einen Monat nach dem Eingriff **Impfstoff gespritzt**, der aus Zellen ihres Tumors und aus einem Stoff, der das Immunsystem stimuliert, besteht. Die Tumorzellen waren vorher mit Strahlen behandelt worden, um sie unschädlich zu machen.

Zum Erstaunen der Ärzte zeigten fast alle Patienten eine Reaktion des Immunsystems. **Nach vier Jahren lebten von den 32 Versuchspersonen noch 94 Prozent.** Bei einer Kontrollgruppe, die nicht geimpft worden war, waren **noch 77 Prozent am Leben.** Durch die Injektion traten auch weniger Zweittumore auf als bei der Kontrollgruppe.

		4y Survival	
		yes	no
Vaccination	yes (n=32)	30	2
	no (n=30)	23	7

Does vaccination yield any effect?

Is the effect “significant”?

Colon carcinoma test

Null hypothesis H_0 :

Vaccination has not (either positive or negative) impact on the patients. The survival rates in the vaccine and non-vaccine group in the whole population are the same.

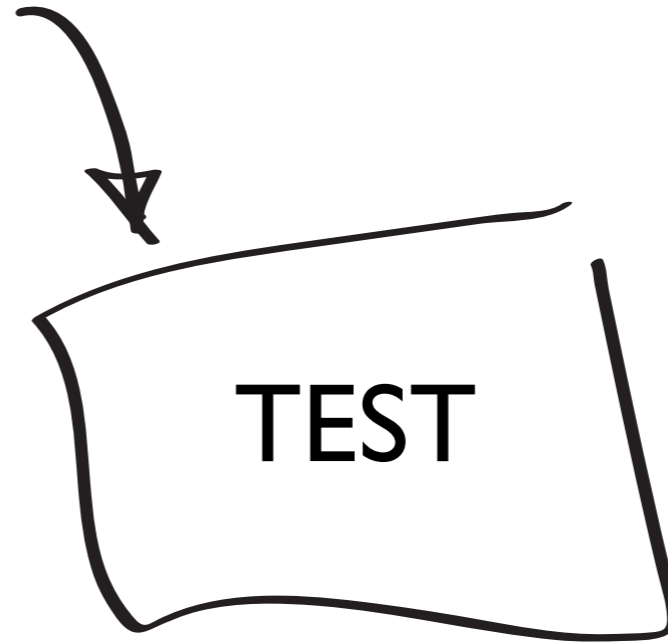
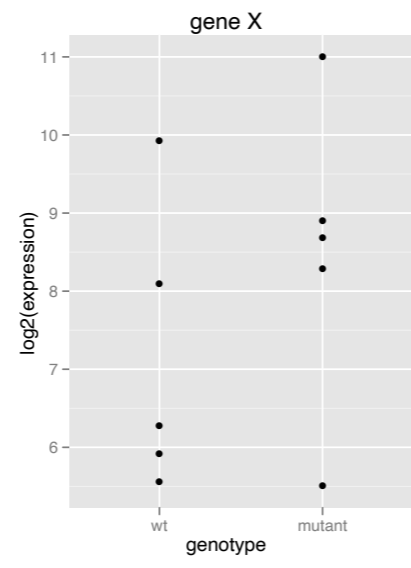
Alternative hypothesis H_1 :

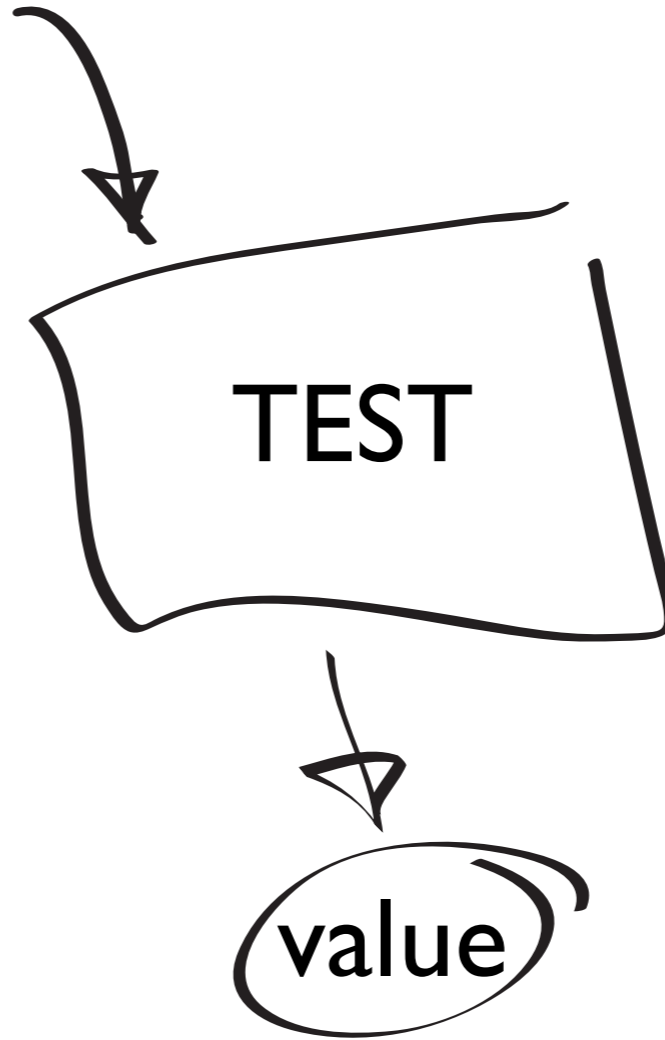
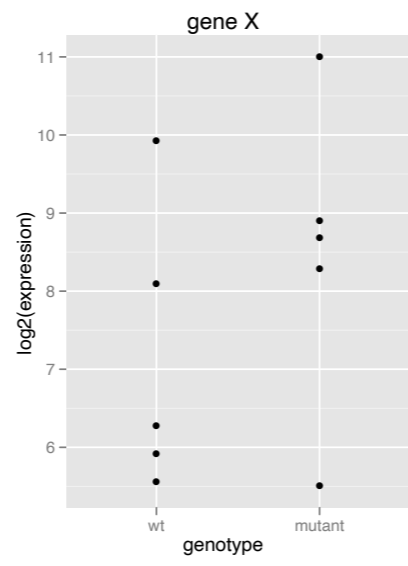
For the whole population, the survival rates in the vaccine and non vaccine group are different.

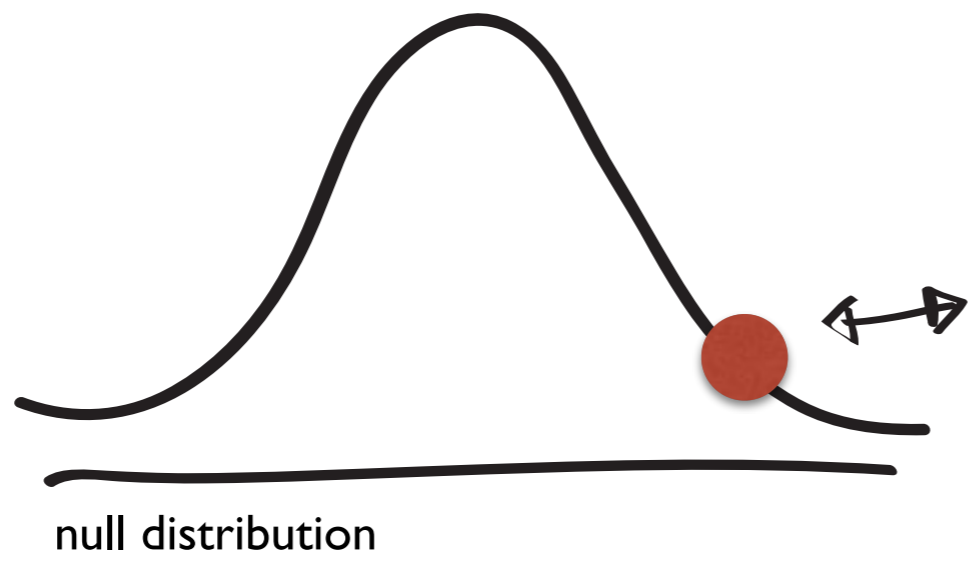
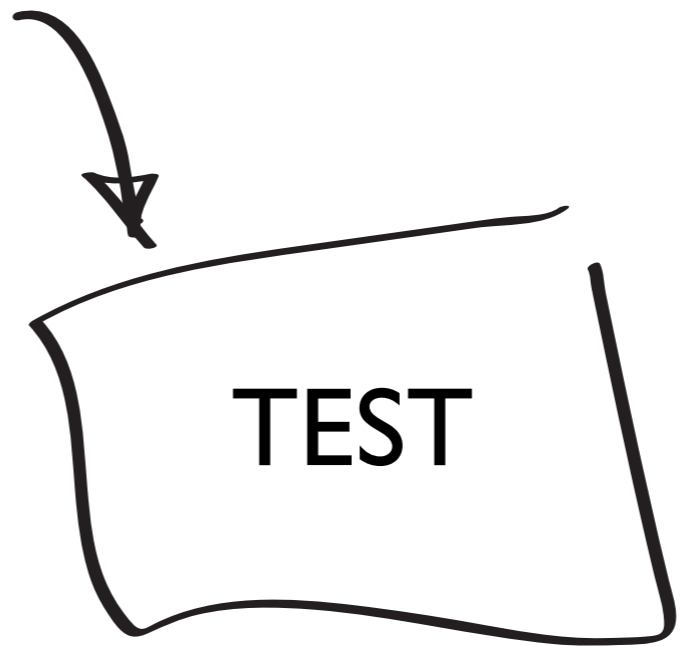
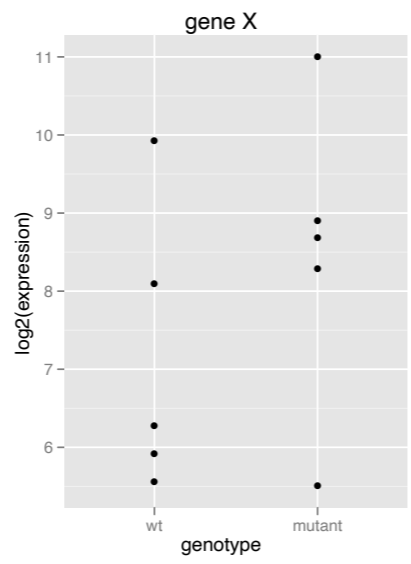
Choose the **significance level α**
(usually: $\alpha = 1\%$; 0.1% ; 5%)

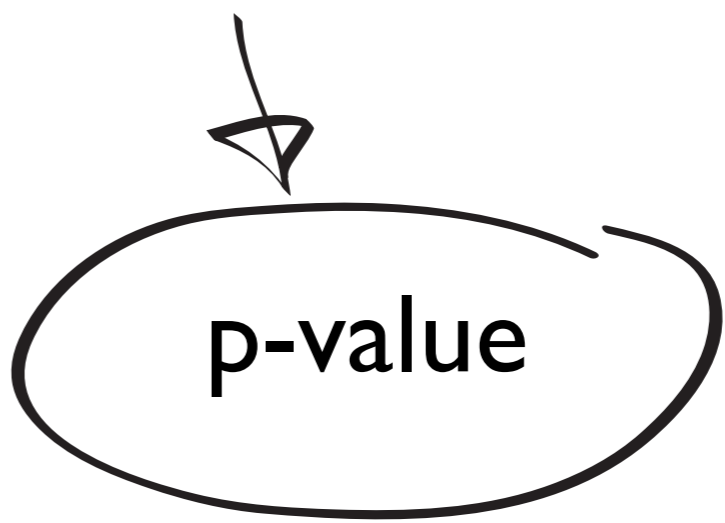
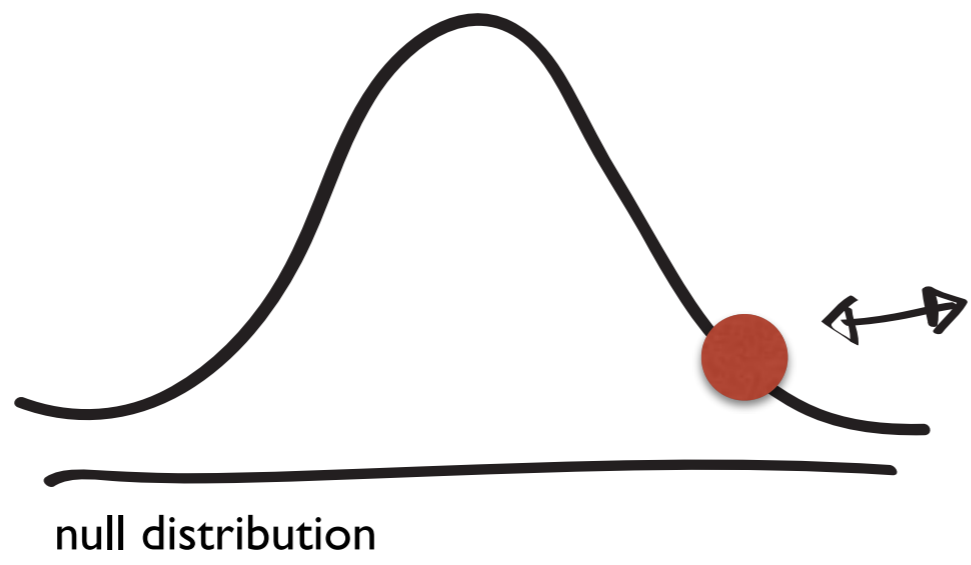
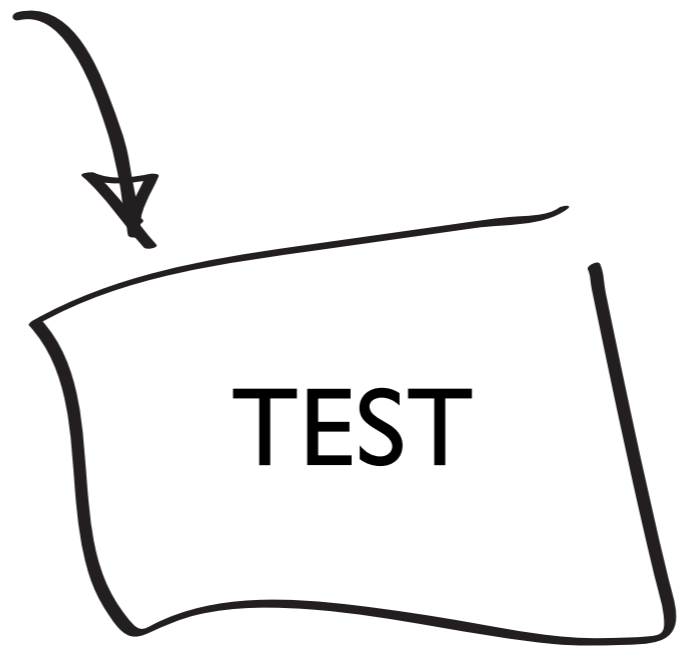
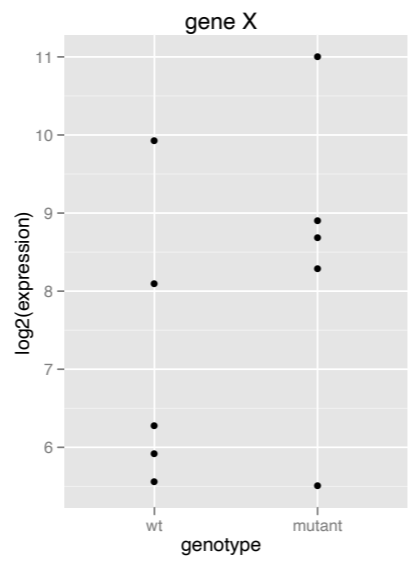
Interpretation of the significance level α :

If there is no difference between the groups, one obtains a false positive result with a probability of α .

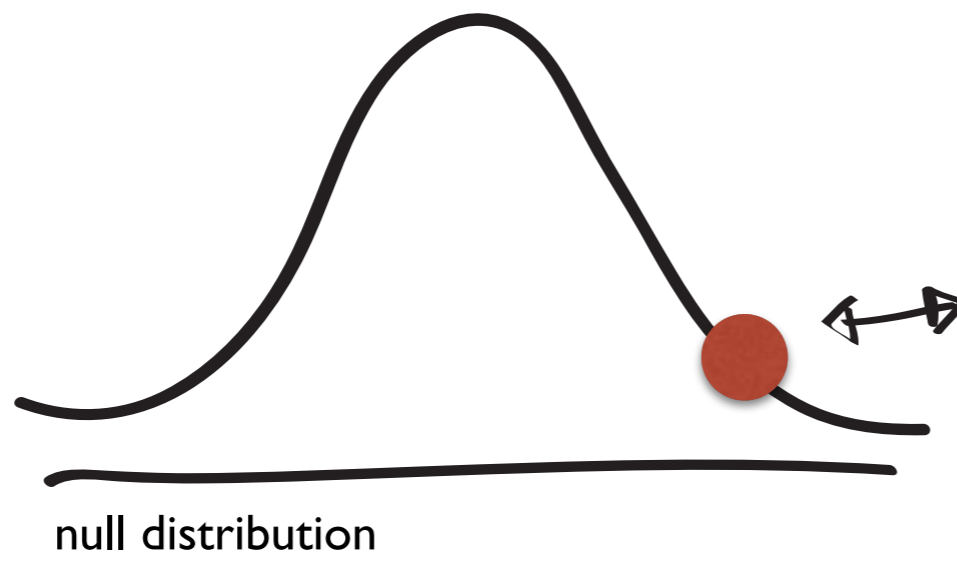
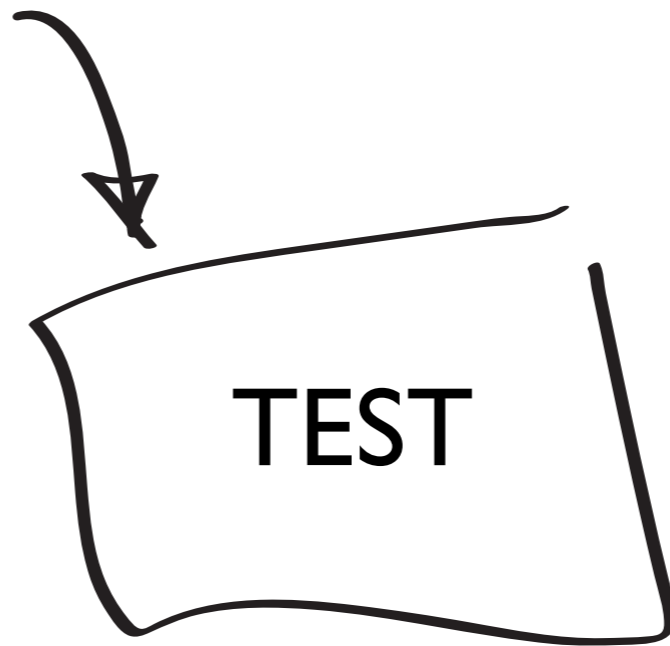
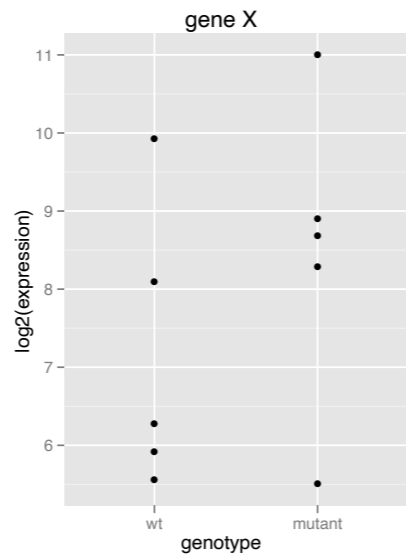






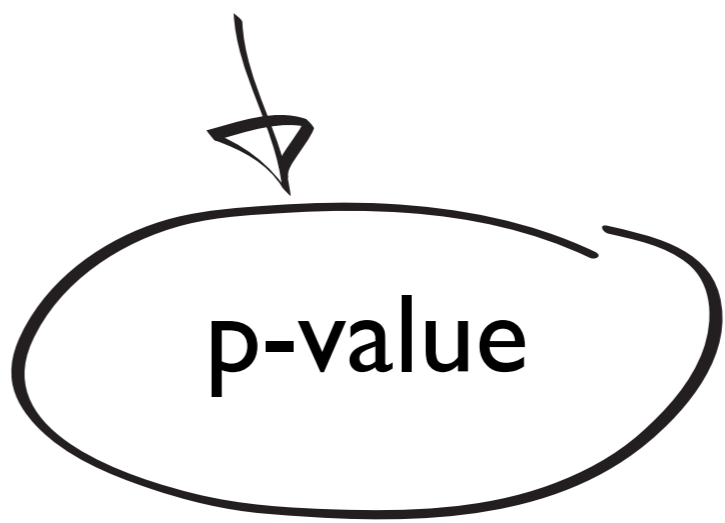
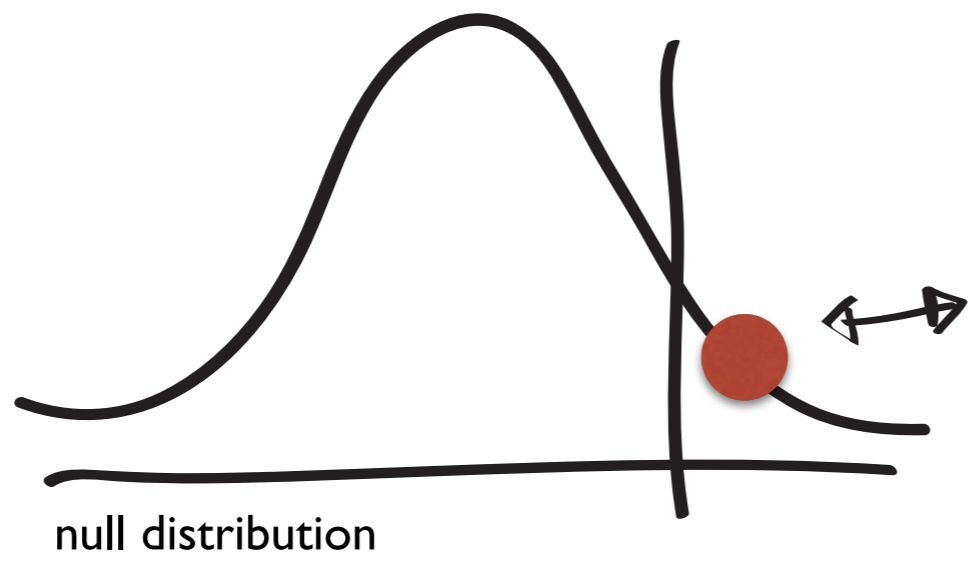
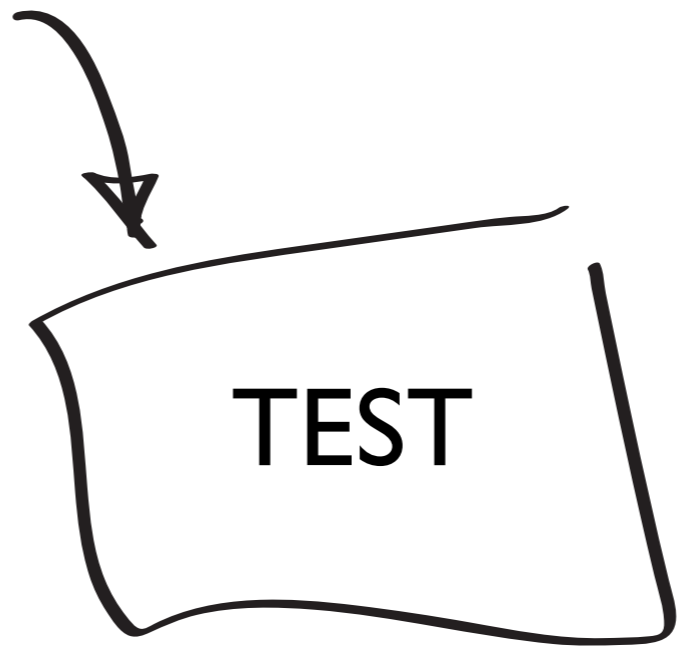
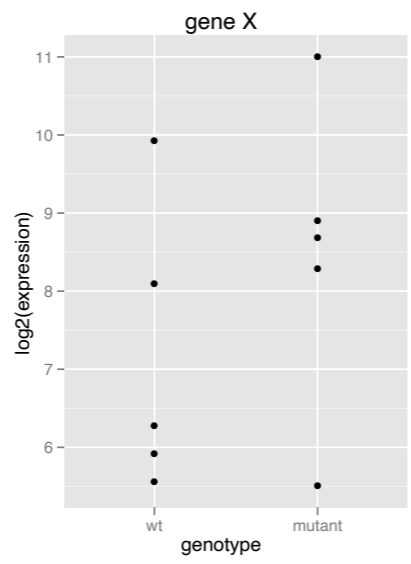


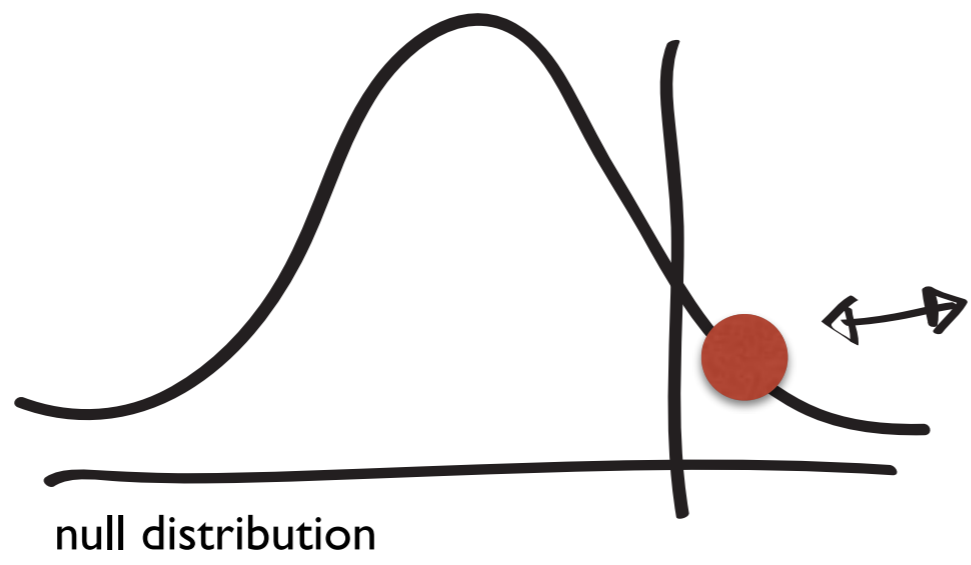
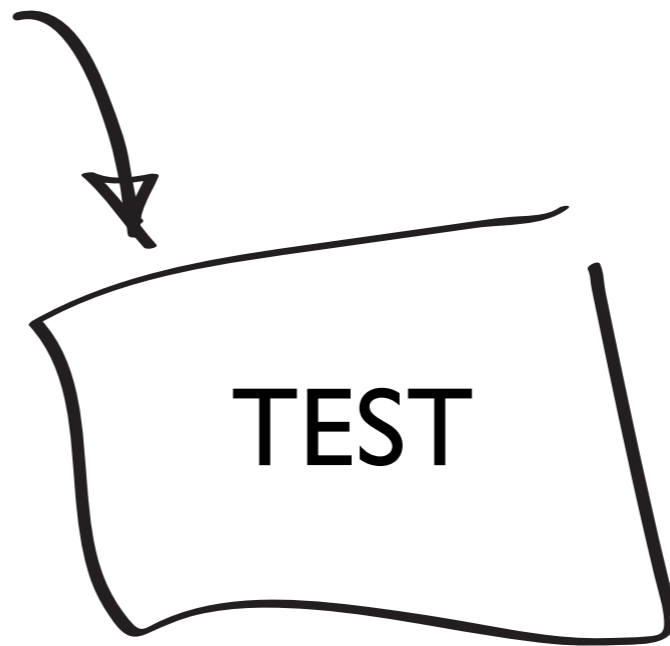
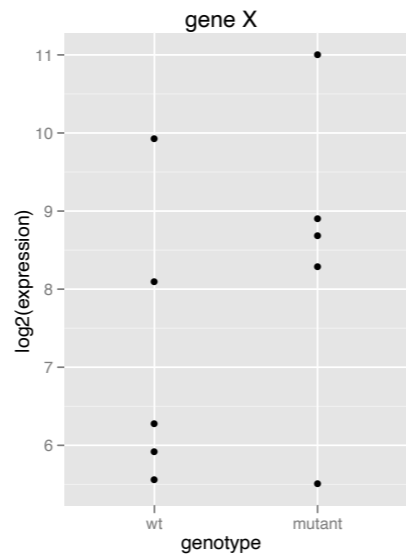
the p-value is the probability to observe an effect of the measured size (or larger) by chance (there was no effect in first place)



value

p-value





value

p-value

significant!!

if the p-value is lower than a pre-defined threshold (α , 0.05) the null hypothesis (no-effect) is rejected and the alternative hypothesis (effect) applies

α also defines the rate of accepting a false positive rejection of the null hypothesis (i.e. 5% false positives, type I error)

Colon carcinoma test

choice of **test statistic**

“Fisher’s Exact Test”



Colon carcinoma test

Value of the test statistic t after the experiment has been carried out. This value can be converted into a p-value:

$$p = 0.0766 \quad 7.7\%$$

Since we have chosen a significance level $\alpha = 5\%$, and $p > \alpha$, we cannot reject the null hypothesis, thus we keep it.

Formulation of the result: At a 5% significance level (and using Fisher's Exact Test), no significant effect of vaccination on survival could be detected.

Consequence: We are not (yet) sufficiently convinced of the utility of this therapy.

But this does not mean that there is no difference at all!

Common Tests

Which test?

- depends on the question asked
- depends on the number of independent (causes) and dependent (effect) variables
- depends on the number of levels of independent variables
- depends on the data type (continuous, discrete, categorical)
- depends on the requirements/assumptions of the test

common assumptions for common tests

- sampling has to be independent
- sample has to be representative of the population

comparing two groups

the experiment

- independent variable: treatment (2 levels)
- dependent variable: a measurement of e.g. enzyme activity, protein level, RNA...
- 5 biological replicates and 3 technical replicate measurements each

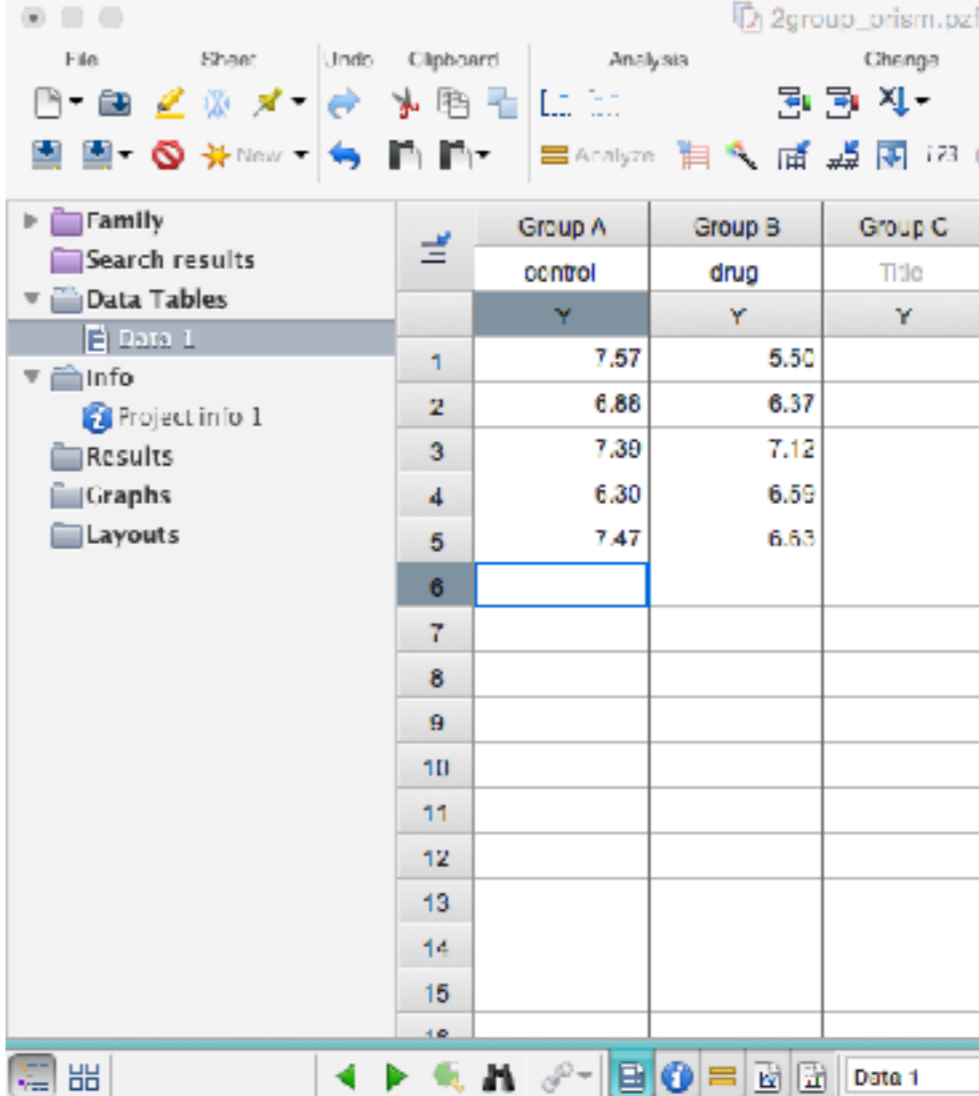
value	treatment	bio.replicate	tech.replicate
7,47	control	1	1
7,19	control	1	2
8,06	control	1	3
6,74	control	2	1
7,49	control	2	2
6,41	control	2	3
7,37	control	3	1
7,23	control	3	2
7,56	control	3	3
6,64	control	4	1
6,14	control	4	2
6,11	control	4	3
7,62	control	5	1
7,69	control	5	2
7,11	control	5	3
5,22	drug	1	1
5,49	drug	1	2
5,79	drug	1	3
6,08	drug	2	1
6,56	drug	2	2
6,47	drug	2	3
6,84	drug	3	1
6,93	drug	3	2
7,58	drug	3	3
6,97	drug	4	1
6,51	drug	4	2
6,28	drug	4	3
6,26	drug	5	1
6,66	drug	5	2
6,98	drug	5	3

the experiment

- first average the technical replicates (e.g. using EXCEL or a calculator)

value	treatment	bio.replicate
7,57	control	1
6,88	control	2
7,39	control	3
6,30	control	4
7,47	control	5
5,50	drug	1
6,37	drug	2
7,12	drug	3
6,59	drug	4
6,63	drug	5

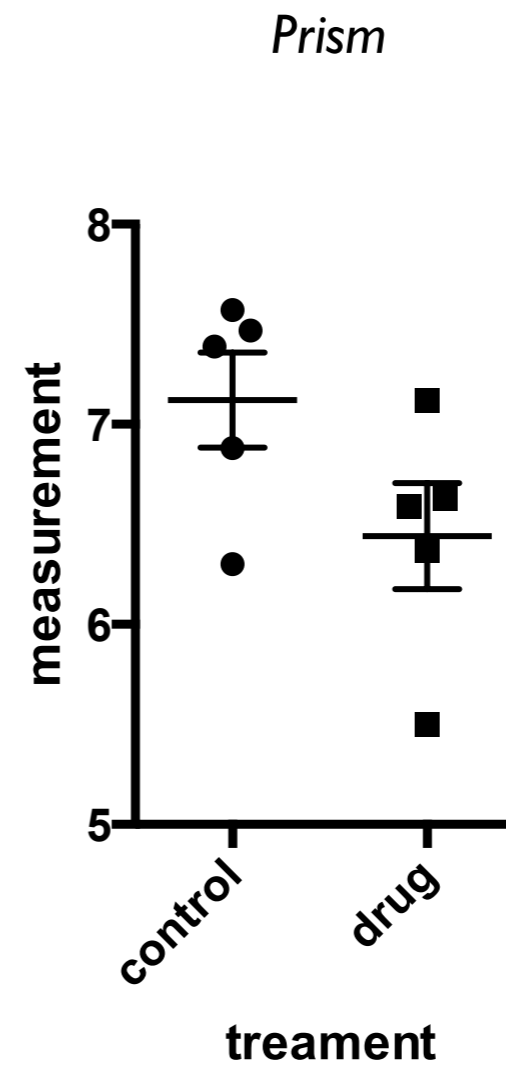
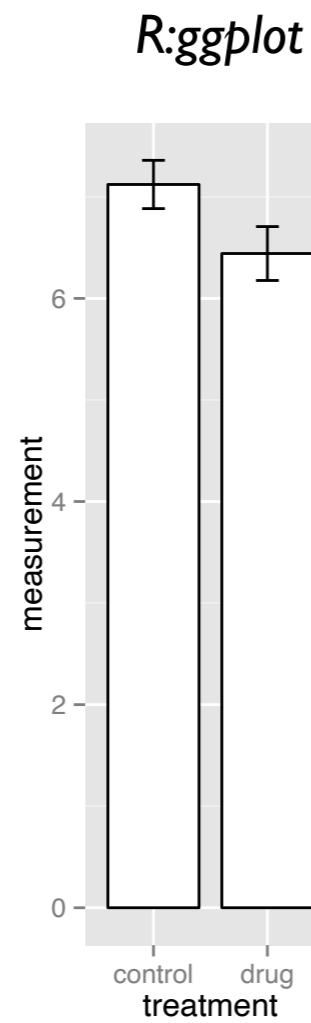
Prism



The screenshot shows the Prism software interface with a data table. The table has columns for Group A, Group B, and Group C, and rows for individual replicates. The data is as follows:

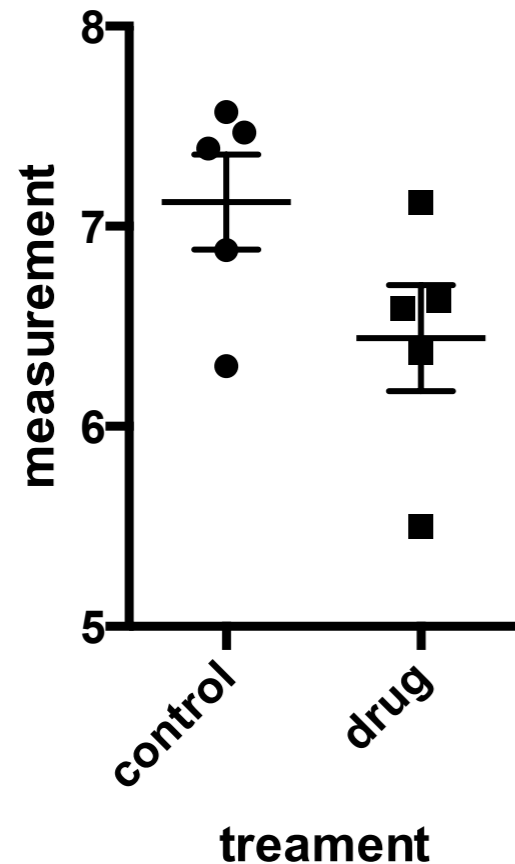
	Group A	Group B	Group C
	control	drug	Title
	Y	Y	Y
1	7.57	5.50	
2	6.88	6.37	
3	7.39	7.12	
4	6.30	6.59	
5	7.47	6.63	
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			

visualise the data



SEM error bars

Two group comparisons



- does the drug have an effect?
- is there a “significant” difference in the measurements?
- null hypothesis:
there is no difference in group means
- alternative hypothesis:
there is a difference in group means
- how likely is such a group means difference occurring by chance?

two sample t-test (unpaired, two-tailed)

requirements:

- a) roughly *equal variances* in both groups
- b) approx. *normally distributed* values
- c) group *sizes can be different*
- d) samples were obtained *independently*

is my data homoscedastic (equal variance)?

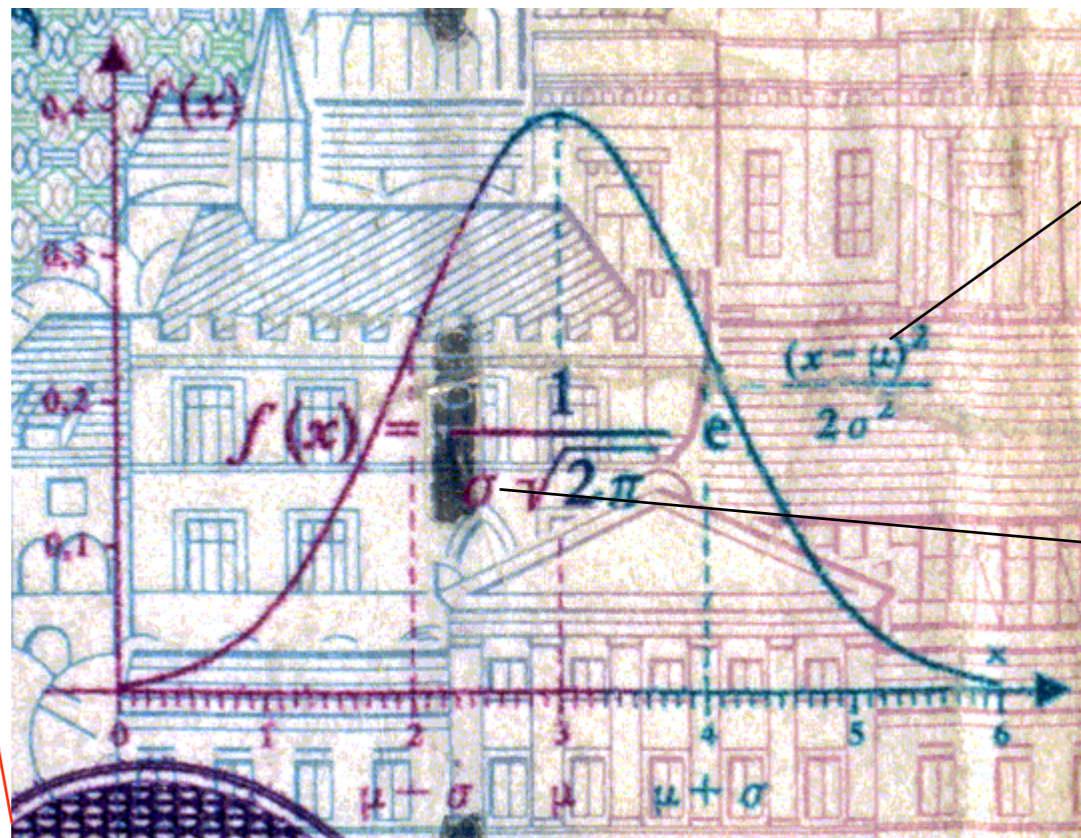
- the measurements can be tested for equal variances (F-test)
- however, the test is very sensitive ...
- and for a small number of n an estimation is not possible
- *pragmatic solution*: always use the t-test with Welch correction which allows for unequal variances

normal distribution

C.F Gauss (1777-1855):

Roughly speaking, continuous variables that are the (additive) result of a lot of other random variables follow a Gaussian distribution. (central limit theorem)

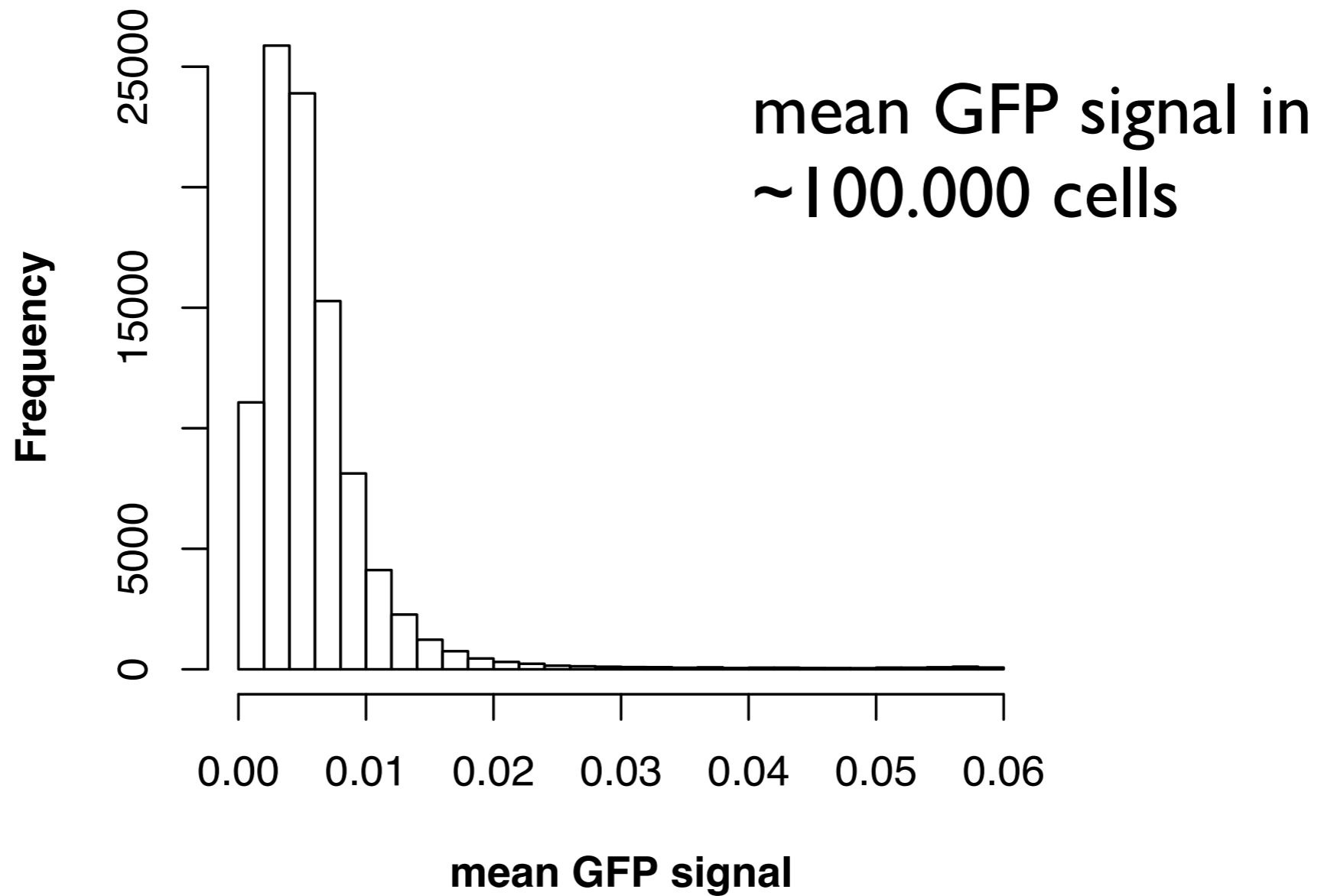
It is often sensible to assume a gaussian distribution for continuous variables.



□ □ □ □ □ □ □ □ □ □ □ □
□ □ □ □ □ □

□ □ □ □ □ □ □ □ □ □
□ □ □ □ □ □ □ □ □ □

non-normal distribution

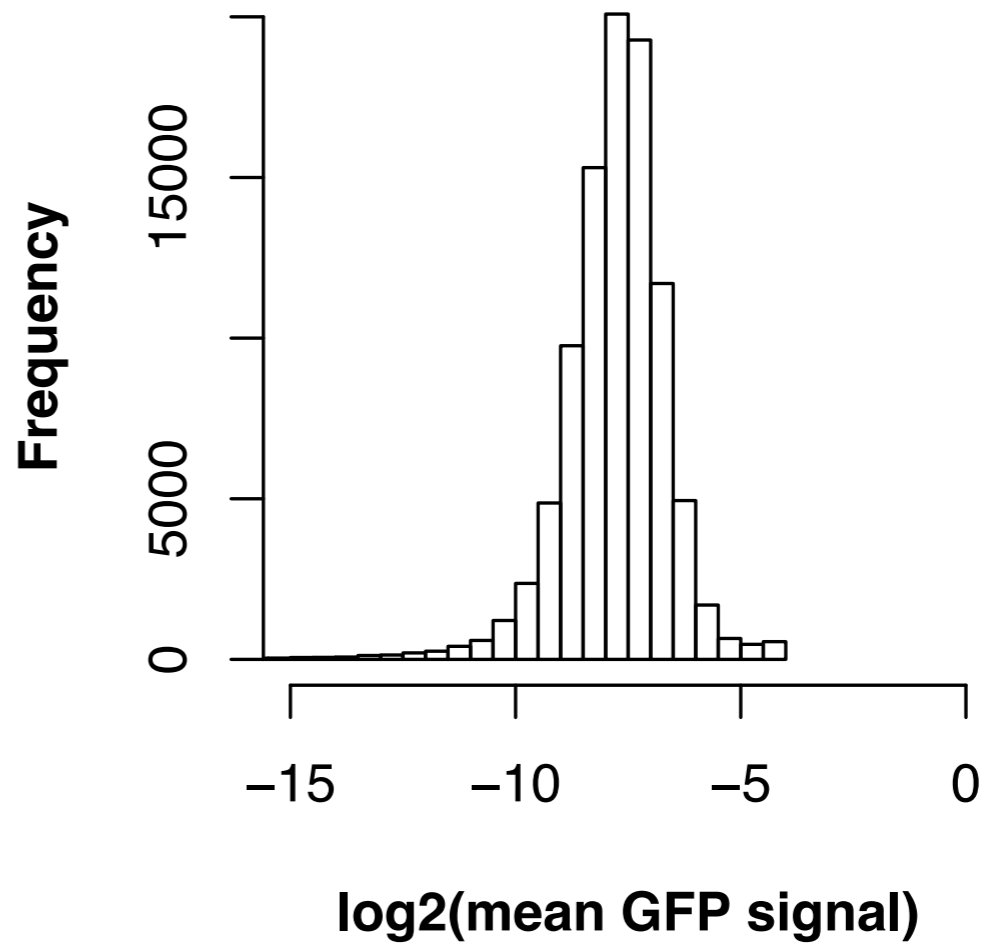
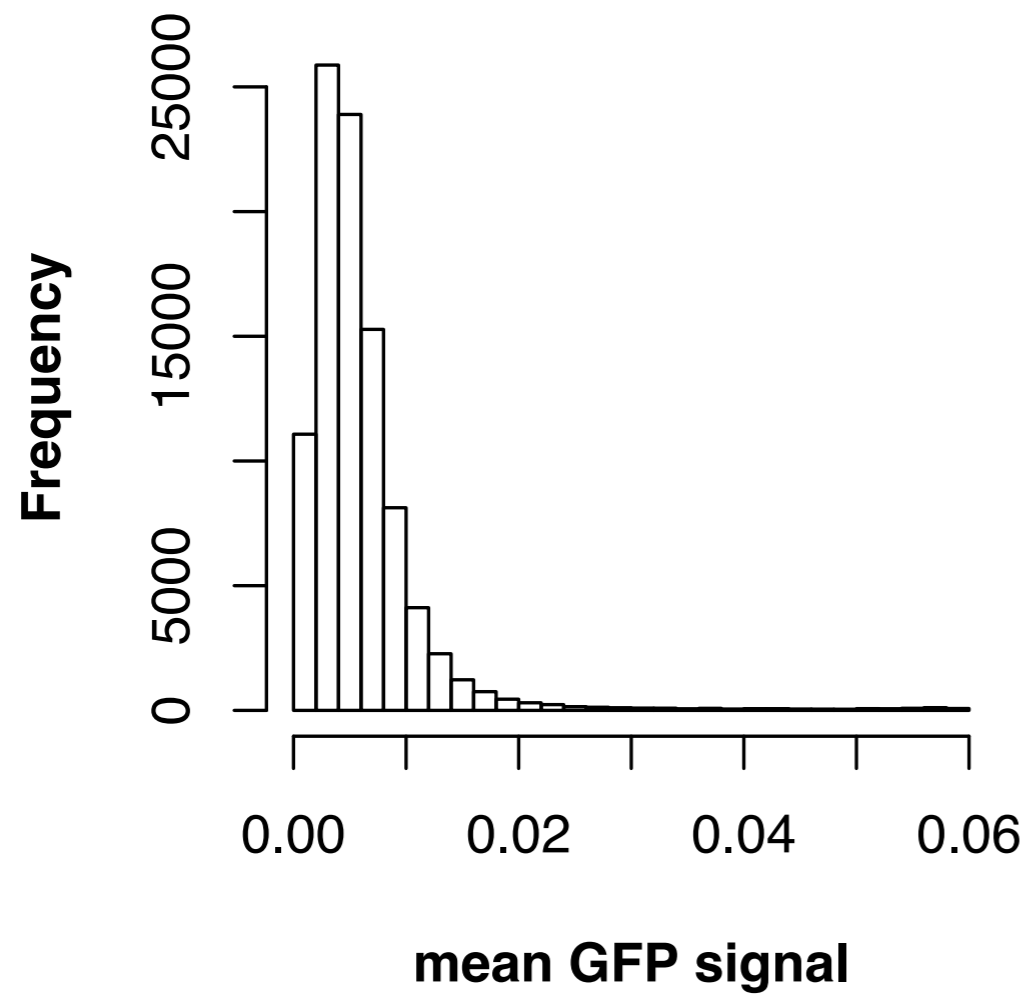


is my data normal?

- Look at a histogram if you sampled sufficiently enough data points ($n \gg 20$). Roughly normal is sufficient.
- If many data points are sampled a formal statistical test can be applied to test for normality (e.g. Shapiro test). However, many data sets that are significantly non-normal would be perfectly appropriate for a t-test or ANOVA.
- The distribution of the population is the important one (not the one of the sample). One therefore might look at other data, too.
- Try data transformation to achieve normality. Gene expression and fluorescence intensity measurements are e.g. known to be normally distributed after log-transformation.

Data transformation!

log-normal distribution



others (rarely to be 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 mycoplasma.

t-test EXCEL

	A	B	C	D	E	F	G
1		control	drug				
2	1	7.57	5.50				
3	2	6.88	6.37				
4	3	7.39	7.12				
5	4	6.30	6.59				
6	5	7.47	6.63				
7							
8		0.09					
9							
10							

X T.TEST function

Returns the probability that is associated with a Student's t-Test. Use **T.TEST** to determine whether two samples are likely to have come from the same two underlying populations that have the same mean.

Syntax

T.TEST(array1,array2,tails,type)

Argument	Description	Remarks
array1	The first data set.	<ul style="list-style-type: none">• None.
array2	The second data set.	<ul style="list-style-type: none">• None.
tails	Specifies the number of distribution tails.	<ul style="list-style-type: none">• If tails = 1, T.TEST uses the one-tailed distribution. If tails = 2, T.TEST uses the two-tailed distribution.• If tails is any value other than 1 or 2, this function returns the #NUM! error value.• If this argument is nonnumeric, this function returns the #VALUE! error value.• If this argument contains a decimal value, this function ignores the numbers to the right side of the decimal point.
type	The kind of t-Test to perform.	<ul style="list-style-type: none">• If type equals 1, T.TEST performs a paired test.• If type equals 2, T.TEST performs a two-sample equal variance (homoscedastic) test.• If type equals 3, T.TEST performs a two-sample unequal variance (heteroscedastic) test.• If this argument is nonnumeric, this function returns the #VALUE! error value.• If this argument contains a decimal value, this function ignores the numbers to the right side of the decimal point.

t-test Prism

Analyze Data

Build-in analysis: ▼

Which analysis?

- Transform, Normalize...
 - Transform
 - Normalize
 - Prune rows
 - Remove baseline and column math
 - Transpose X and Y
 - Fraction of Total
- XY analyses
- Column analyses
 - t tests (and nonparametric tests)
 - One-way ANOVA (and nonparametric tests)
 - Column statistics
 - Frequency distribution
 - ROC Curve
 - Bland-Altman method comparison
 - Correlation
 - Identify outliers
- Grouped analyses
- Contingency table analyses
- Survival analyses
- Parts of whole analyses
- Generate curve

Analyze which data sets?

- A: control
- B: drug

Parameters | Tests (and Nonparametric Tests)

Experimental Design | Options

Experimental design

- Unpaired
- Paired

	Group A	Group B
	Control	Treated
	Y	Y
1		
2		
3		
4		
5		

Assume Gaussian distribution?

- Yes. Use parametric test.
- No. Use nonparametric test.

Choose test

- Unpaired t test. Assume both populations have the same SD
- Unpaired t test with Welch's correction. Do not assume equal SDs

Cancel OK

Unpaired t test

	Table Analyzed	2 Group
1	Table Analyzed	2 Group
2		
3	Column B	drug
4	vs.	vs.
5		
6	Column A	control
7	Unpaired t test with Welch's correction	
8	P value	0.0931
9	P value summary	ns
10	Significantly different? (P < 0.05)	No
11	One- or two-tailed? P value?	Two-tailed
12	Welch-corrected t, df	t=1.908 df=7.903
13		
14	How big is the difference?	
15	Mean ± SEM of column A	7.122 ± 0.2375, n=5
16	Mean ± SEM of column B	6.442 ± 0.2656, n=6
17	Difference between means	-0.6800 ± 0.3562
18	95% confidence interval	-1.503 to 0.1431
19	R squared	0.3166
20		
21	Fixed to compare variances	
22	F, DFn, Dfd	1.250, 4, 4
23	P value	0.3342
24	P value summary	ns
25	Significantly different? (P < 0.05)	No
26		
27		
28		

RF

Unpaired t test of 2 Group

t-test R

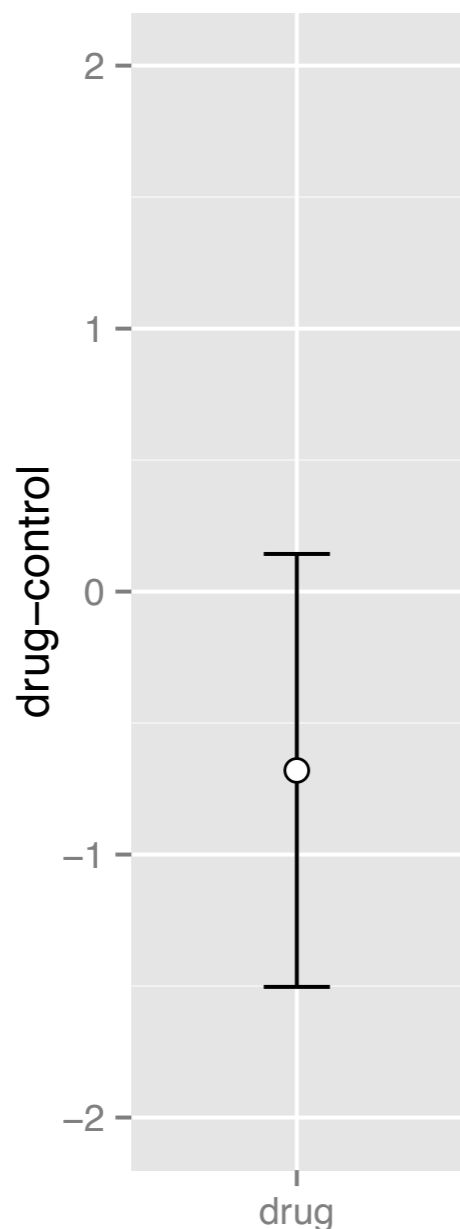
```
> mat
      value treatment bio.replicate
1  7.573993   control             1
2  6.879926   control             2
3  7.387153   control             3
4  6.299270   control             4
5  7.468581   control             5
6  5.500070    drug              1
7  6.371208    drug              2
8  7.118872    drug              3
9  6.585543    drug              4
10 6.633153    drug              5

> t.test(value~treatment,data=mat)

Welch Two Sample t-test

data:  value by treatment
t = -1.909, df = 7.906, p-value = 0.09311
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.5031541  0.1431231
sample estimates:
 mean in group drug mean in group control
          6.441769          7.121785
```

CI of group mean difference



- 95% CI of the group mean difference ranges from -1.5 to 0.14
- spans 0, i.e. includes no difference
- provides a measure of the effect size and significance!
- better than p-value!!

another example

Data was collected to test whether treating cultured cells with a drug increases the activity of an enzyme. Five different clones of the cell were tested. With each clone, control and treated cells were tested side by side.

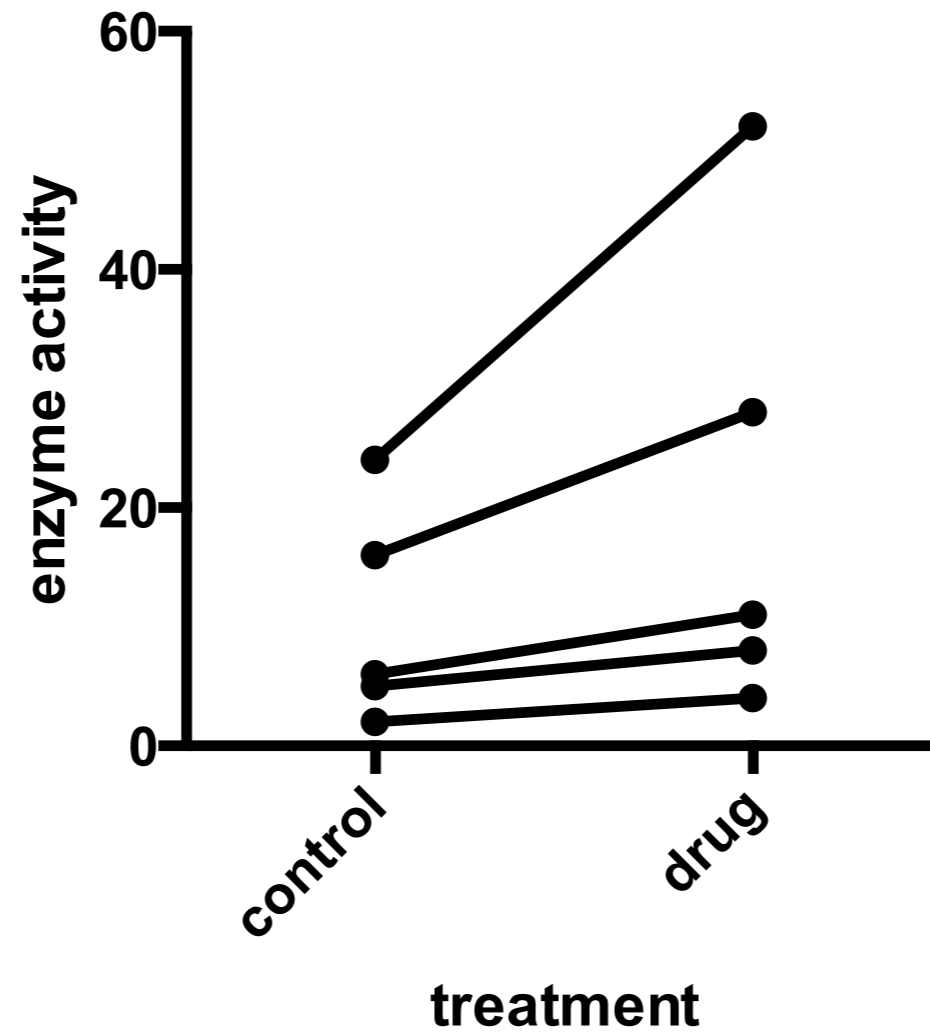
control	treated	
24	52	← clone 1
6	11	← clone 2
16	28	← clone 3
5	8	← clone 4
2	4	← clone 5

t-test

control	treated	
24	52	28
6	11	5
16	28	12
5	8	3
2	4	2

$p=0.107$ (t-test, unpaired, two-tailed)

graphical representation



the ratio t-test (one sample t-test)

the ratio is much more informative (biologically)
but the ratio is asymmetric: log transformation! (here $\log 10$)

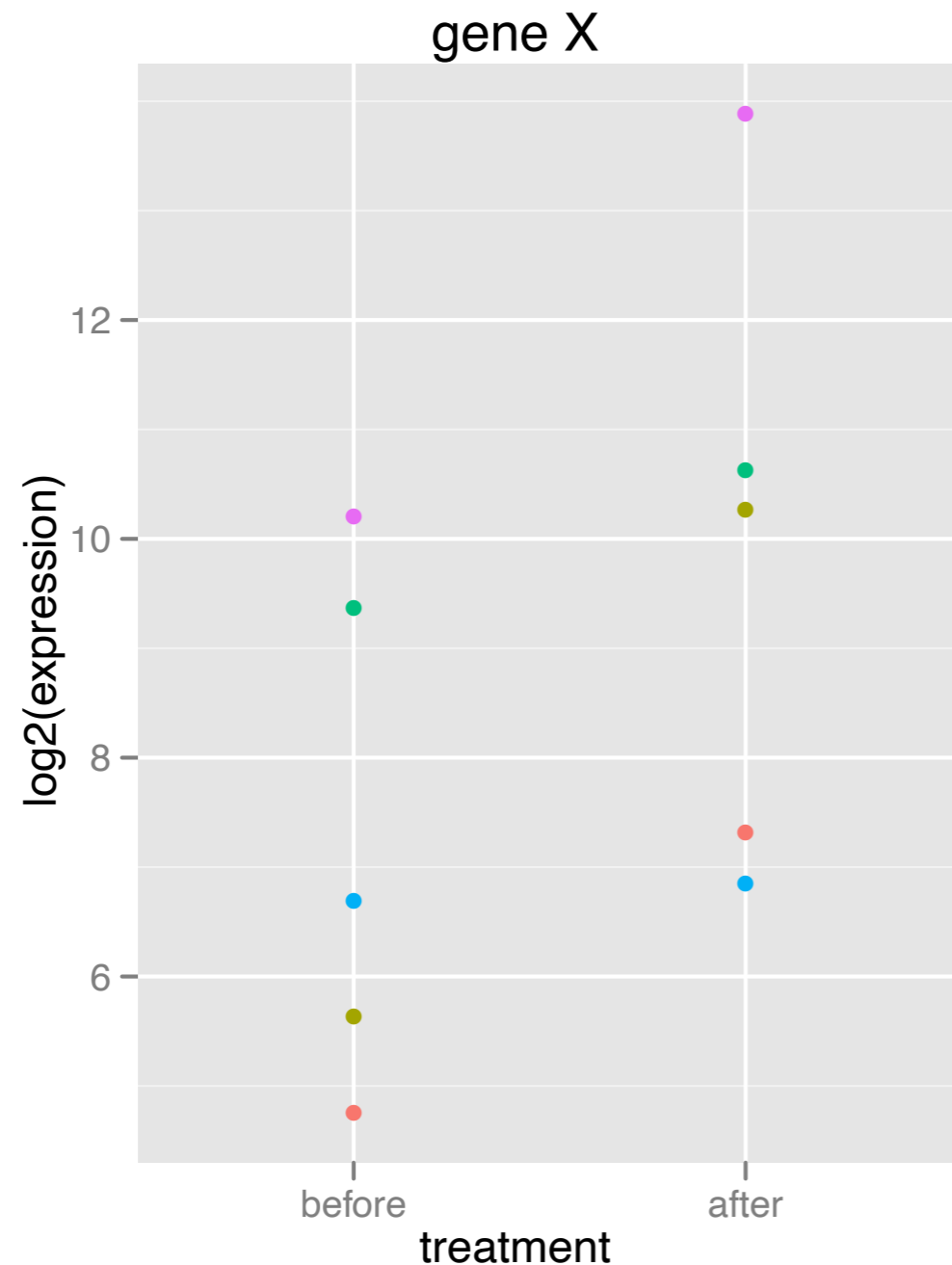
control	treated	log ratio
1,38	1,72	0,34
0,78	1,04	0,26
1,20	1,45	0,24
0,70	0,90	0,20
0,30	0,60	0,30

p-value = 0.0003 (t-test, $\mu=0$, two-tailed)

mean change: 0.26 (1.86 antilogged)

CI 95%: 0.20-0.33 (1.61-2.15 antilogged)

Two group comparisons, paired data



is there a “significant”
difference in expression?

two sample t-test (paired, two-tailed)

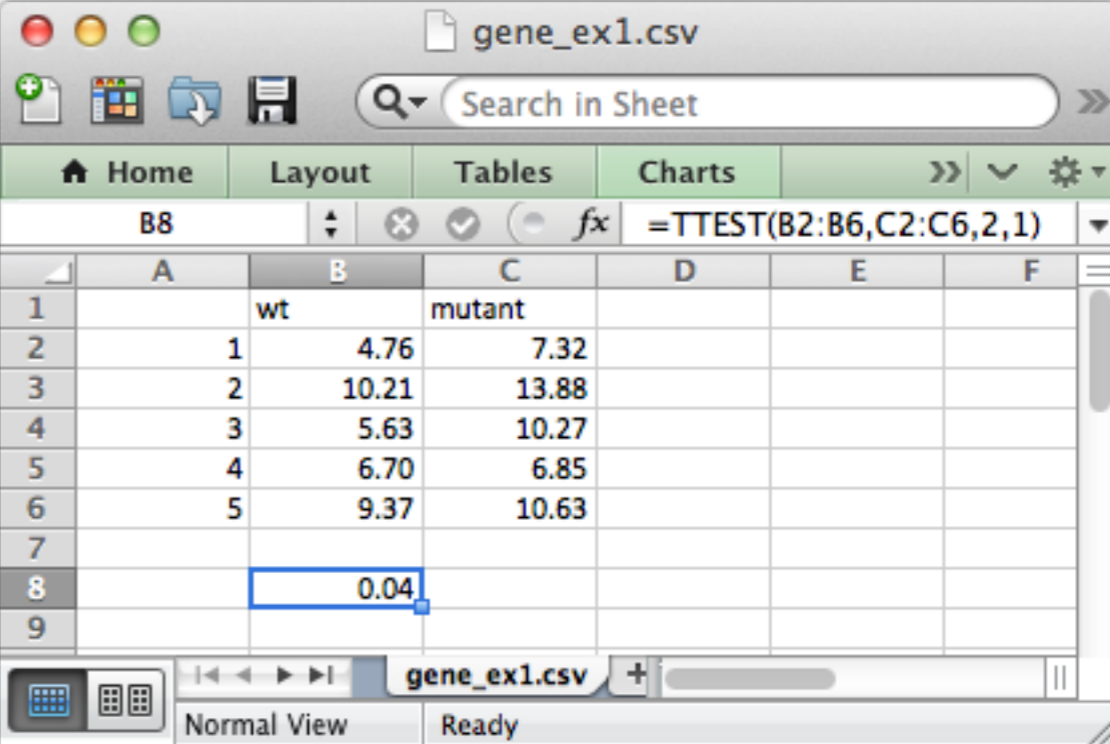
requirements:

- approx. *normally distributed* values
- paired data

```
> t.test(expression.level~treatment, data=df, paired=T)
```

```
Paired t-test
```

```
data: expression.level by treatment  
t = -3.0556, df = 4, p-value = 0.03782  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -4.6915433 -0.2245871  
sample estimates:  
mean of the differences  
 -2.458065
```



	A	B	C	D	E	F
1		wt	mutant			
2	1	4.76	7.32			
3	2	10.21	13.88			
4	3	5.63	10.27			
5	4	6.70	6.85			
6	5	9.37	10.63			
7						
8		0.04				
9						

Paired tests

- use if:
 - you measure a variable in each subject *before and after* an intervention
 - you run a laboratory experiment several times, each time
 - with a control and treatment preparation handled in *parallel*
 - whenever the value of one subject in the first group is expected to be more similar to particular subject in the the second group than to a random subject in the second group
- The statistical power of a paired test in a paired experimental layout is much higher than for an unpaired test in a paired layout.
- the decision about pairing has to be made before collecting the data!

the ratio t-test versus paired test

control	treated	log ratio
1,38	1,72	0,34
0,78	1,04	0,26
1,20	1,45	0,24
0,70	0,90	0,20
0,30	0,60	0,30

p-value = 0.0003 (t-test, **paired**, two-tailed)

p-value = 0.0003 (t-test, $\mu=0$, two-tailed)

mean change: 0.26 (1.86 antilogged)

CI 95%: 0.20-0.33 (1.61-2.15 antilogged)

one-tailed or two-tailed?

one tailed only if there is absolutely
no possibility for a movement in the
other direction

and

the decision for this test has been
taken *before* data collection

Summary t-test

- Incredibly powerful 2-group comparison test
- Very few formal requirements: normality is most important
- Parameters:
 - paired: crucial
 - unequal variance can be set by default
 - always 2-sided

Power analysis of two-tailed unpaired T-test

- sample size

- effect size

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

pioneer experiments required to get mean difference and s !

- α , significance level (0.05)

- power, $1 - \beta$ (the probability of making a type II error)

(typically set to 80% or 90%)

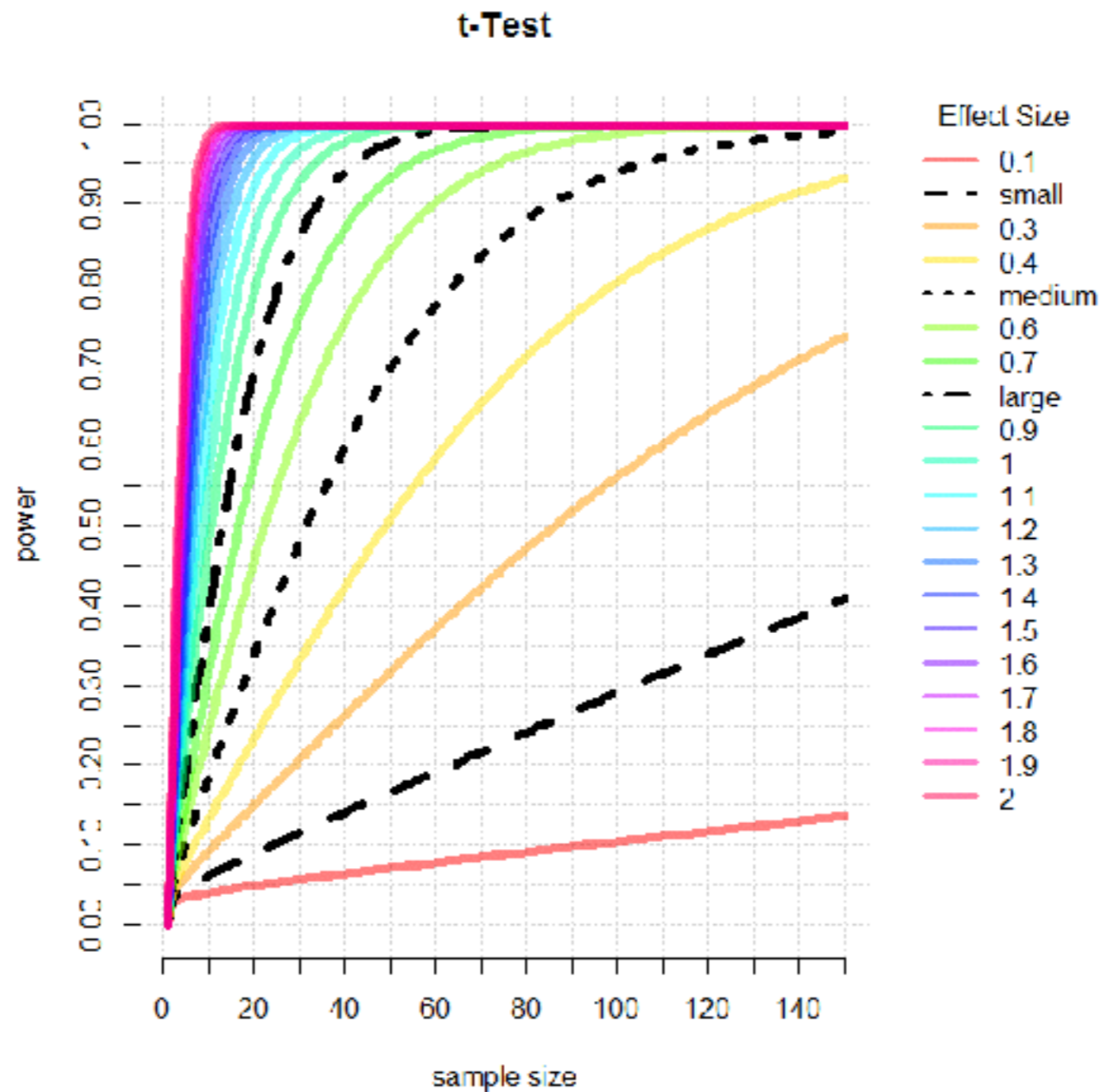
how many samples to detect 2 fold change with a SD of 0.4?

Two-sample t test power calculation

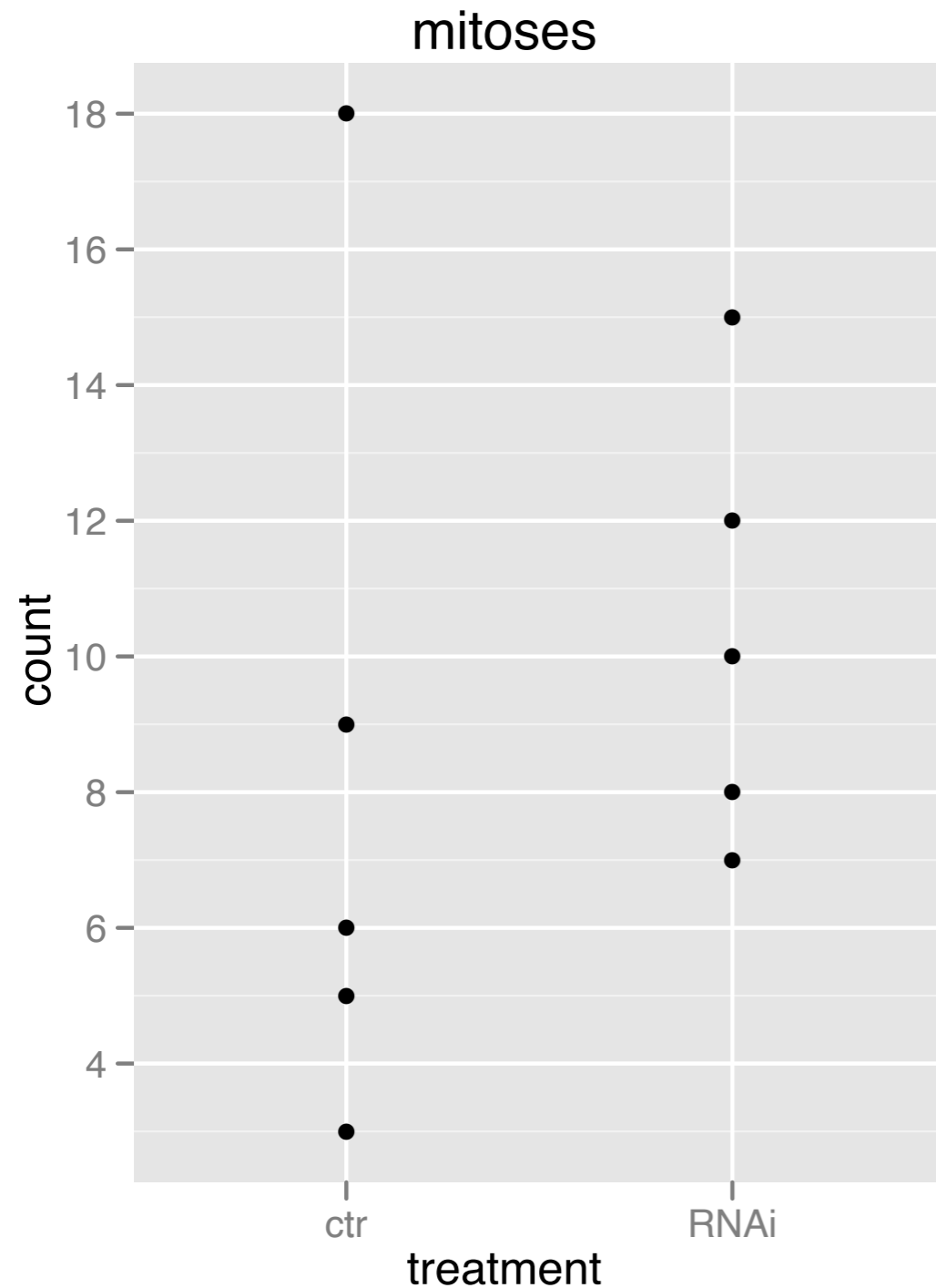
```
n = 4.574784  
d = 2.5  
sig.level = 0.05  
power = 0.9  
alternative = two.sided
```

NOTE: n is number in *each* group

2-sample t-test unpaired



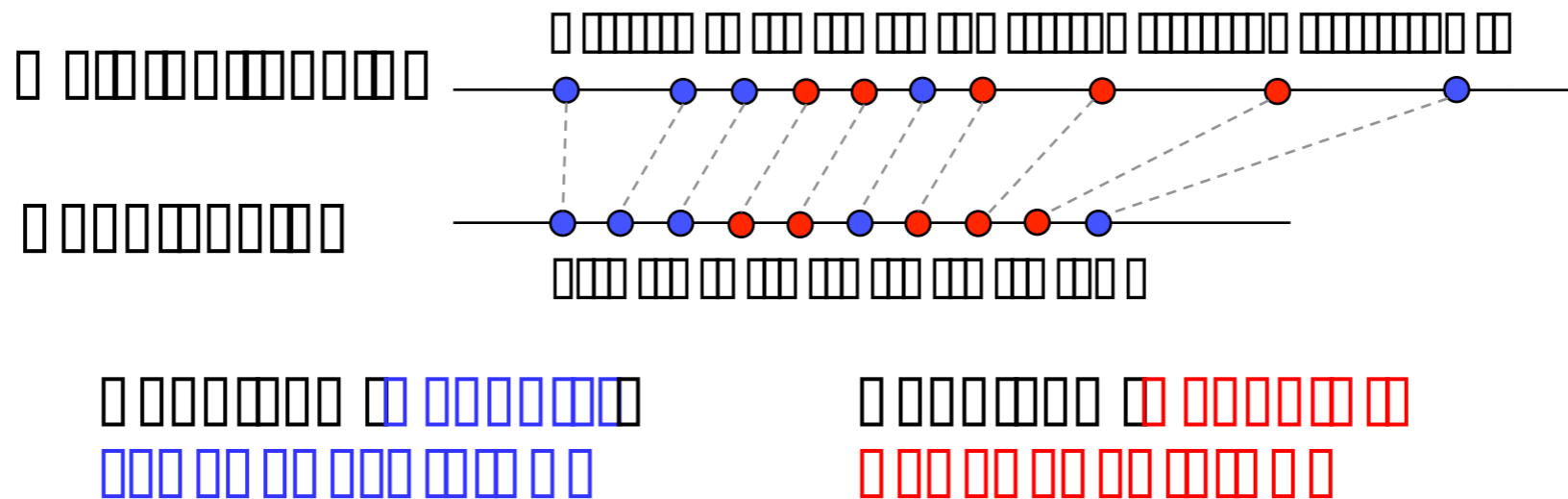
Two group comparisons, non-normal distribution



is there a “significant”
difference in number of
mitotic cells?

Rank Tests (Wilcoxon, Mann-Whitney, U-Test)

	□ □□□□□□				
□□□□	□□□	□□	□□	□□	□□
□□ □□	□□□	□□□	□□	□□	□□□



```
> wilcox.test(mitoses~treatment,data=df)
```

Wilcoxon rank sum test

data: mitoses by treatment

W = 7, p-value = 0.3095

alternative hypothesis: true location shift is not equal to 0

Wrong Test - does it matter?

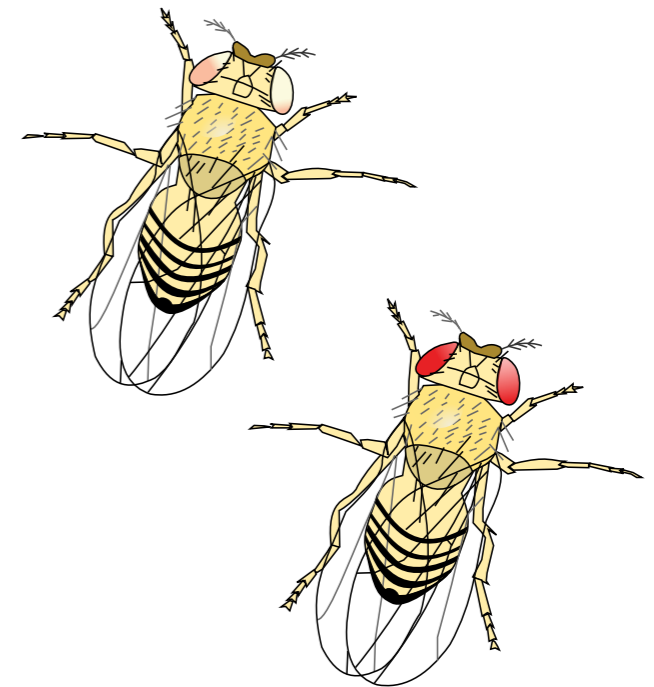
- for large data sets ($n > 50$) a wrong decision does not matter
- for small data sets the wrong choice matters:
 - nonparametric tests have low power
 - parametric tests are not robust

Unpaired binary data

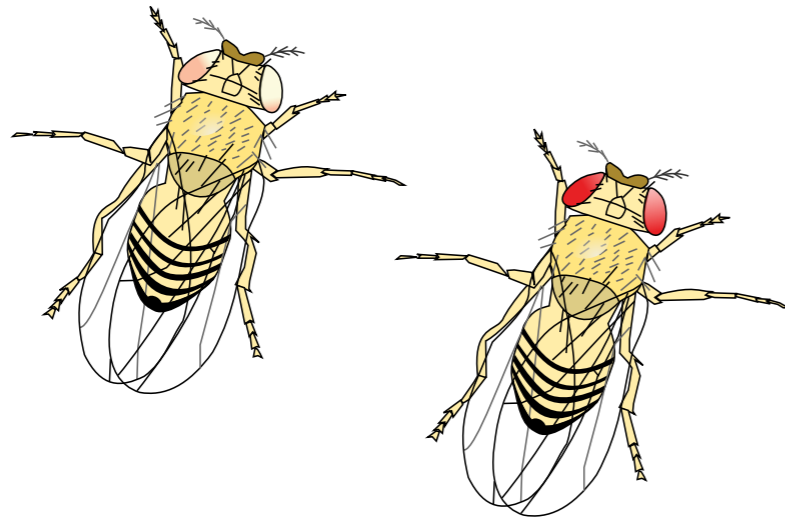
Drosophila embryos are fed with a drug or a control substance. The hatched adults are tested for eye color (either “red” or “white”).

100 flies of 185 treated with drug develop red eyes. 75 flies of 185 treated with a control substance develop red eyes.

Is there a significant effect of the drug?

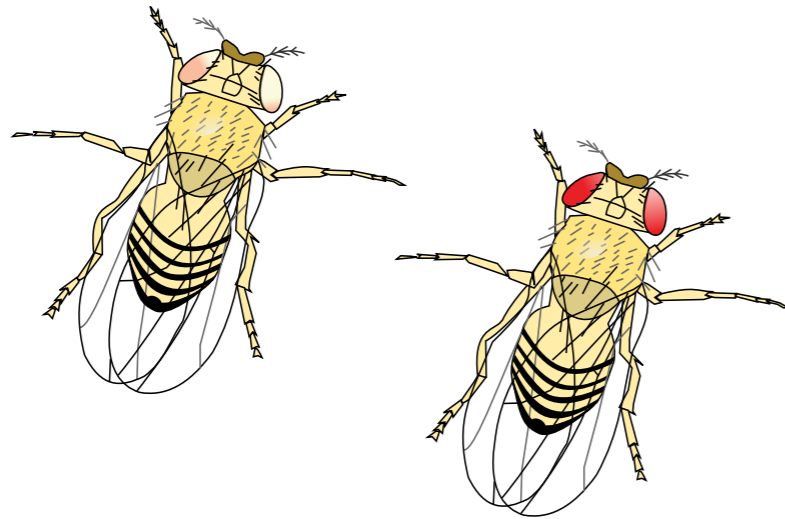


Unpaired binary data



		Red Eyes	
		yes	no
Drug	yes (n=185)	100	85
	no (n=185)	75	110

Unpaired binary data



```
      Eyes
Treatment red white
drug    100    85
ctr      75   110
> fisher.test(FLIES)
```

Fisher's Exact Test for Count Data

```
data:  FLIES
p-value = 0.01235
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.119345 2.661245
sample estimates:
odds ratio
 1.722947
```


**What if we increase the
number of sample
objects?**

Unpaired binary data

```
      Eyes
Treatment  red white
  drug 10000  8500
  ctr   7500 11000
> fisher.test(FLIES)
```

Fisher's Exact Test for Count Data

```
data:  FLIES
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.655480 1.798476
sample estimates:
odds ratio
 1.725494
```

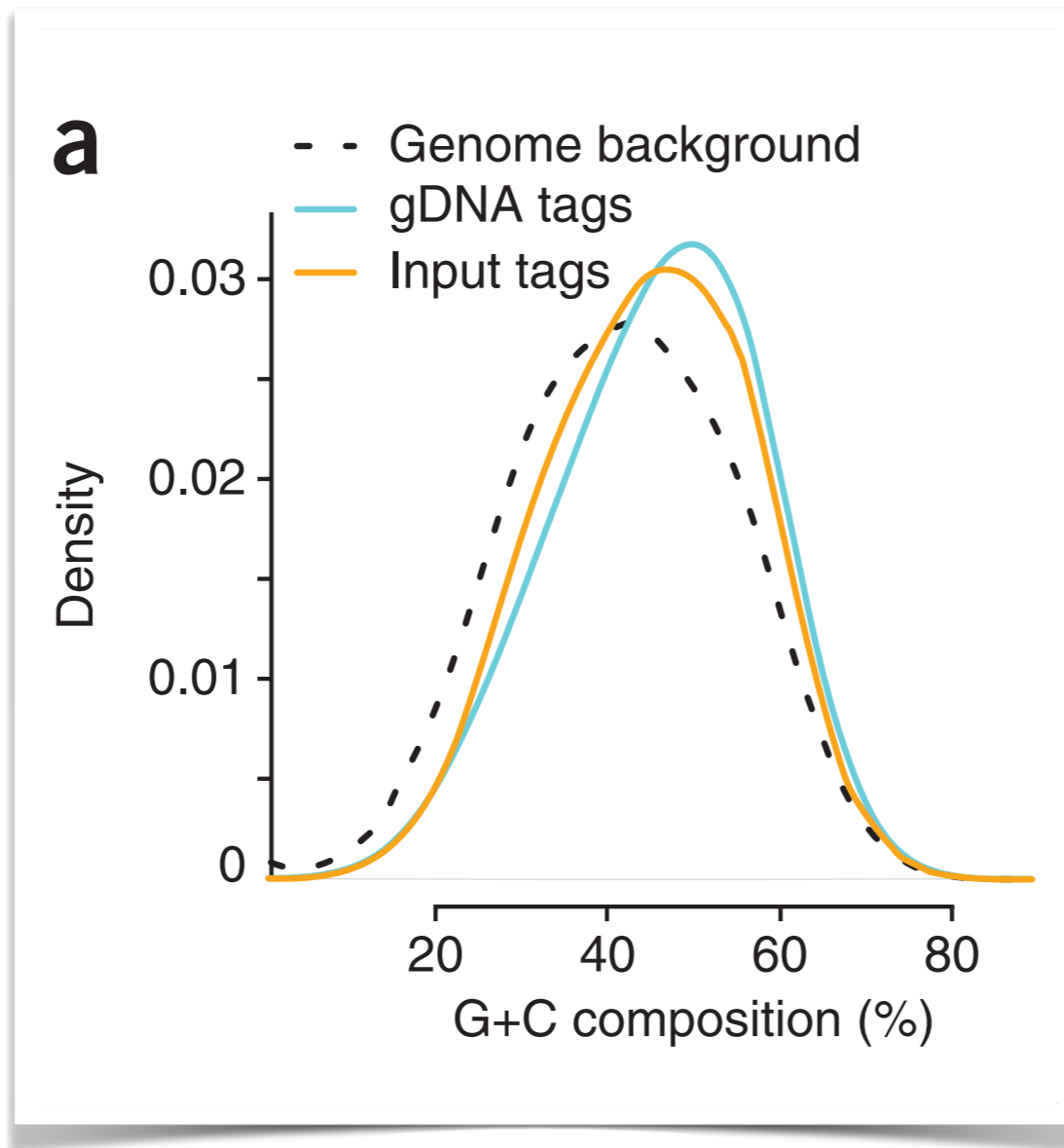
Unpaired binary data

```
      Eyes
Treatment  red white
  drug 10300  9990
  ctr   9700 10010
> fisher.test(FLIES)
```

Fisher's Exact Test for Count Data

```
data:  FLIES
p-value = 0.001999
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.022848 1.106714
sample estimates:
odds ratio
 1.063969
```

high n increases the
probability to reject H_0



genome background (Online Methods and **Fig. 1a**). Sequencing reads from the chromatin input and gDNA samples had different G+C composition distributions (median, 44% and 47%, respectively; Mann-Whitney test, $P < 2.2 \times 10^{-16}$; **Fig. 1a**), suggesting that chromatin may affect sequencing coverage.

We compared the gDNA read count-normalized coverage

reporting p-values

- report the test applied and the test parameters.
- avoid the terms “statistically significant” and variations thereof (“extremely significant”).
- avoid categorisation of p-values ($p < 0.05$, $p < 0.01$..), just report the p-value as computed with 2 decimal precision.
- upon treatment with X we observed an increase in Y (p-value 0.002, Fisher’s exact test, two-sided).
- always report n (*biological replicates*)

a p-value is a p-value

- a p-value is not necessarily a proxy for the robustness of an effect
- many applications produce “technical p-values” which cannot give any information on biological robustness.
Examples: Database searches, peptide identification in mass spectrometry, ChIPSeq peak calling and other *within-experiment* analyses

Problems of p -values

- p -values are only valid if the assumptions of the underlying test are met
- most importantly, the samples have to be independent and representative of the population

Problems of p-values

- Performing multiple tests within an experiment increases the probability to get a false positive result (a “significant” effect)
- e.g. *simultaneous* testing of many endpoints (genes, proteins) in high throughput studies or simultaneous pairwise comparison of many groups or *sequential* testing
- in order to control for the overall type I error rate the p-values have to be adjusted.

Multiple Testing

Examples:

- Simultaneous testing of many endpoints
(e.g. genes in a microarray study)
- Simultaneous pairwise comparison of many (k) groups
(k pairwise tests = $k(k-1)/2$ tests)



Although each individual test keeps the significance level (say $\alpha = 5\%$), the probability of obtaining (at least one) false positive increases dramatically with the number of tests: $\alpha_k = 1 - (1 - \alpha)^k$.

For 6 tests, the probability of a false positive is already $>25\%$!

The expected number of significant results in a series of k independent hypothesis tests when all null hypotheses are actually true is simply calculated as: $k * \alpha$

in a microarray study interrogating 10000 genes the expected number of false positives is 500

multiple testing correction

One possible solution: p-value correction for multiple testing, e.g. **Bonferroni correction**:

Each single test is performed at the level α/m („local significance level α/m “), where m is the number of tests.

The probability of obtaining a (at least one) false positive is then at most α („multiple/global significance level α “)

Ex.: $m = 6$

Desired multiple level: $\alpha = 5\%$

→ local level: $\alpha/m = 5\%/6 = 0.83\%$

Other solutions: Bonferroni-Holm, Benjamini-Hochberg, Control of False discovery rate (FDR) instead of significance at the group level (family wise error rate, FWER)

- ***Bonferroni*** correction (control of the FWER):
FWER= probability of getting at least one false positive.
The critical value (alpha) for an individual test is obtained by dividing the familywise error rate (usually 0.05) by the number of tests.
Thus if you are doing 100 statistical tests, the critical value for an individual test would be $0.05/100=0.0005$, and you would only consider individual tests with $P<0.0005$ to be significant.
- ***Benjamini-Hochberg*** (control of FDR):
controls the proportion of significant results being false positives.

Repeated Testing to Reach Significance

needs adjustment!

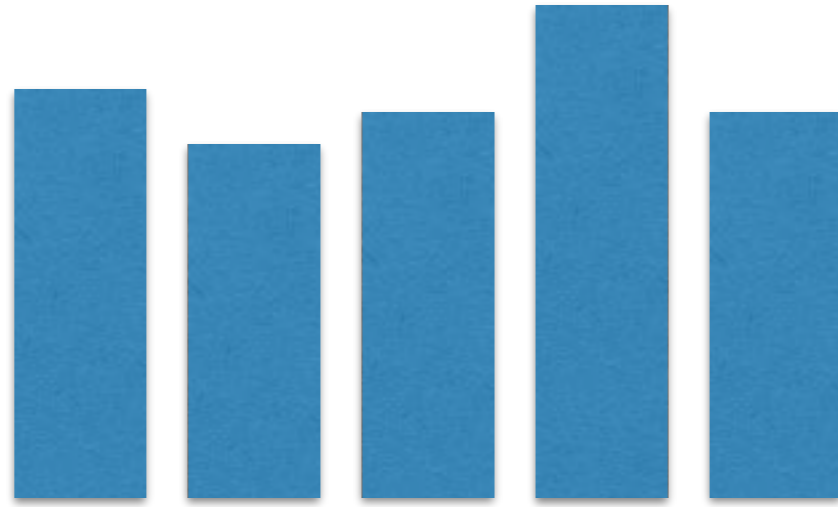
DON'T DO IT

"If you torture your data long enough, they will tell you whatever you want to hear." (Mills ,1993).

p-value hacking (fishing)

Simmons JP, Nelson LD, Simonsohn U. 2011. False-Positive Psychology: **Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.** Psychological Science 22: 1359–1366.

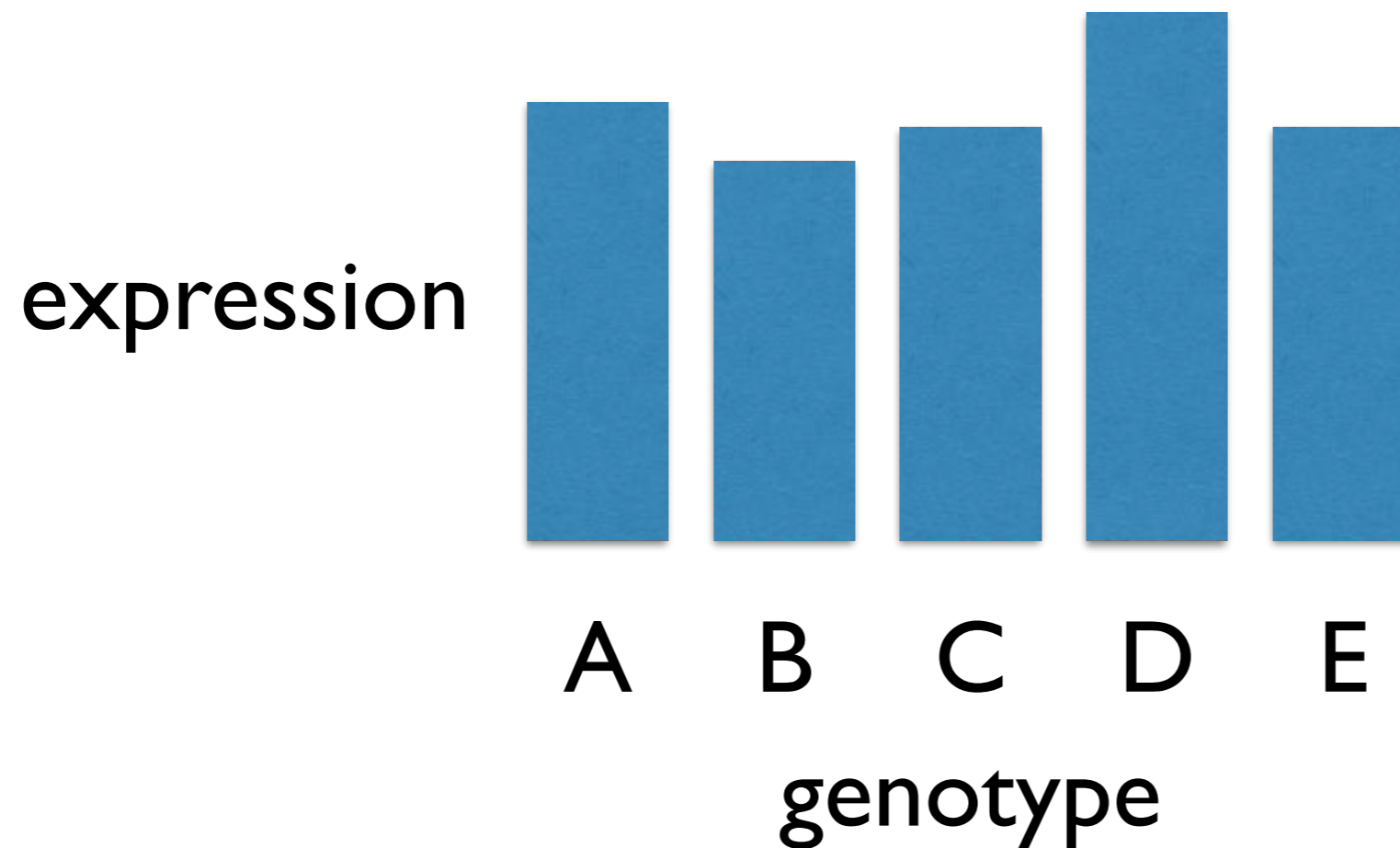
- sampling bias, the “drawer problem”
- trying different testing procedures
- sequential testing
- multiple endpoints reporting only the significant ones



ANOVA

- measure differences in more than 2 groups (avoiding multiple testing corrections when using standard t-tests)
- can be used to analyse the contribution of different sources of variation to a response

example



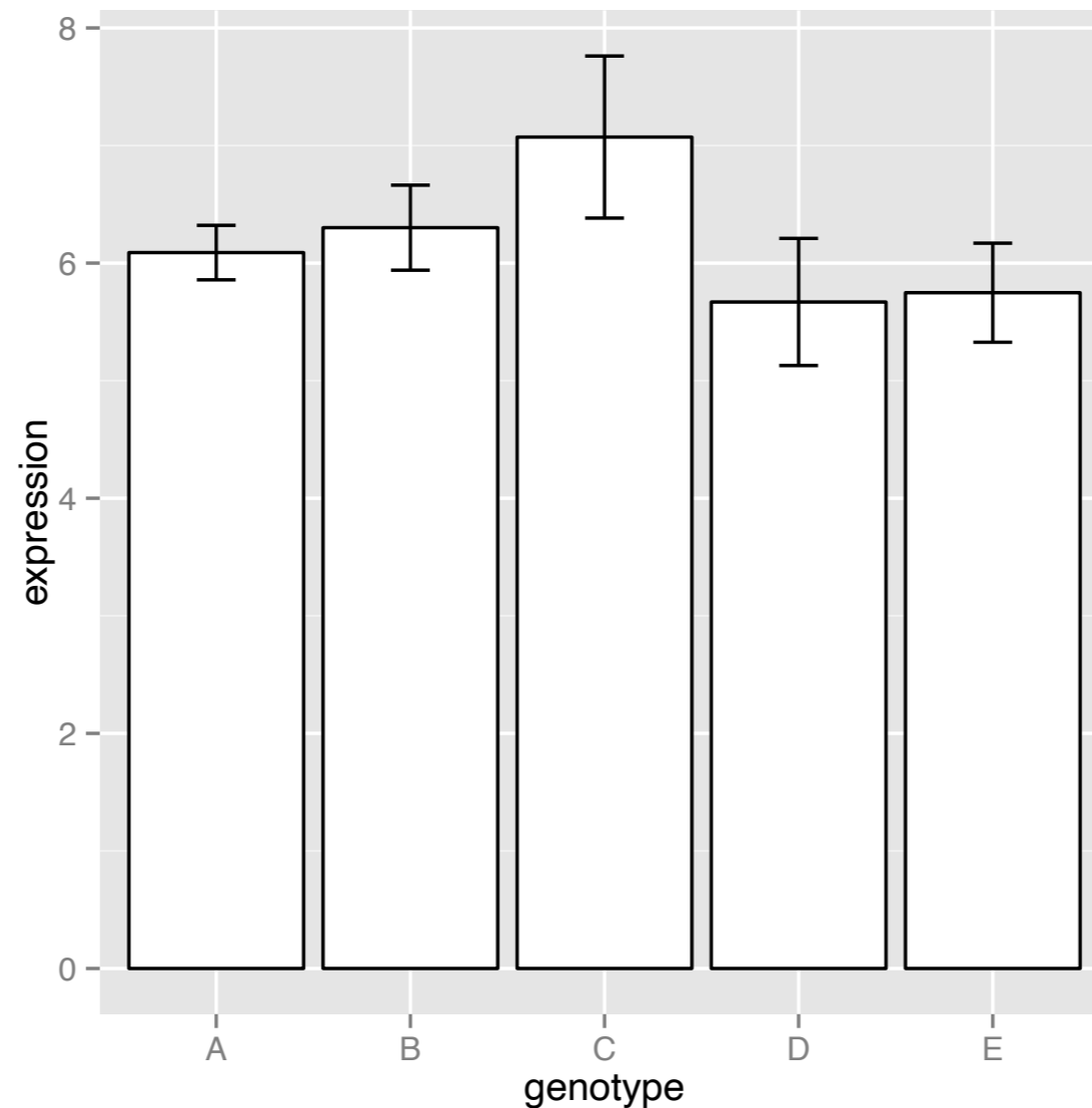
Null hypothesis: means of the measurement variable (expression) are the same for the different categories of data (genotype)

Alternative hypothesis: the means of expression are not all the same

Assumptions to be met

- observations in each group are normally distributed
- standard deviations in the groups should be equal (homoscedastic). this is particularly important in unbalanced designs (unequal number of observations)
- independency, random selection

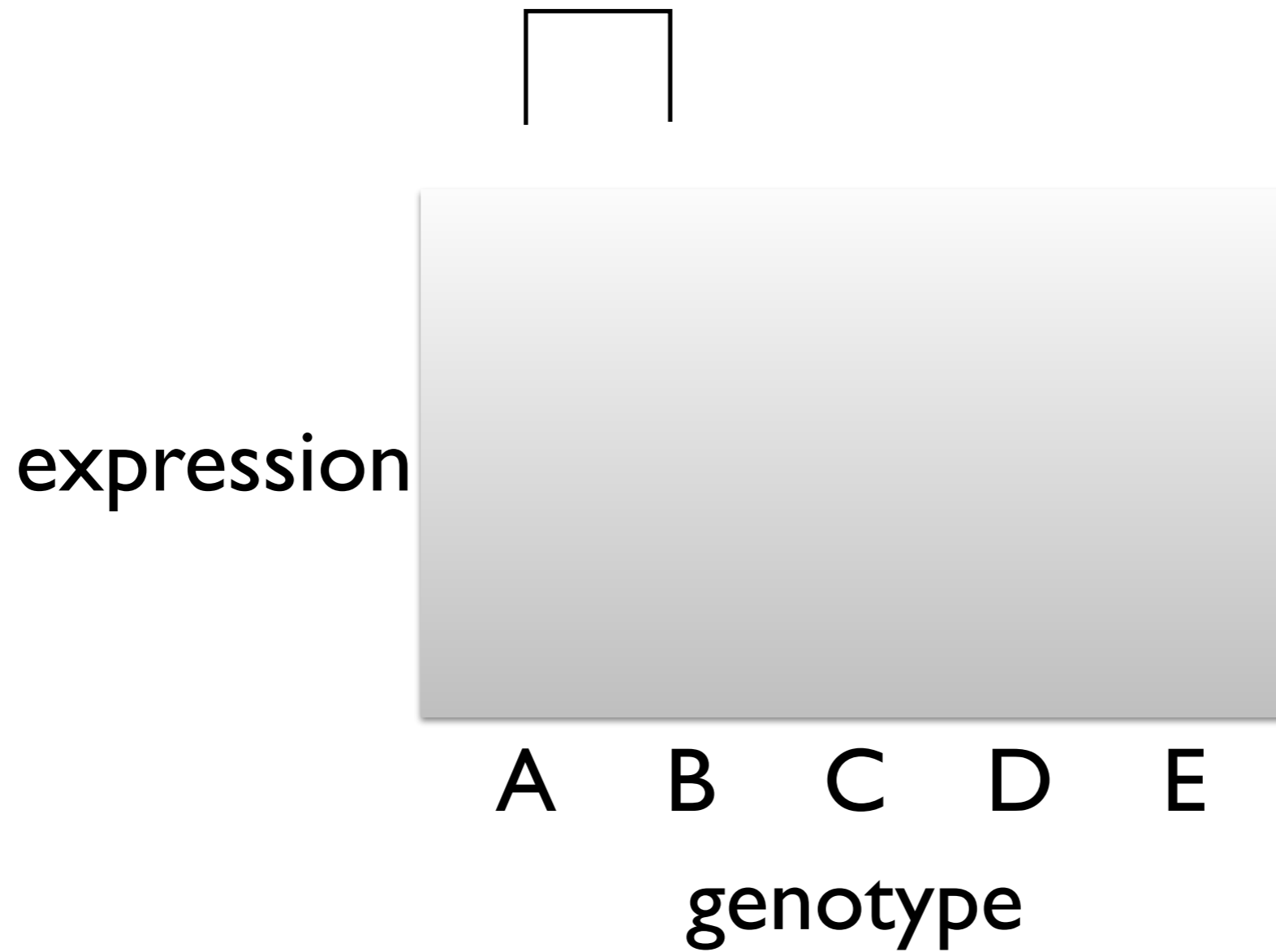
reporting the result



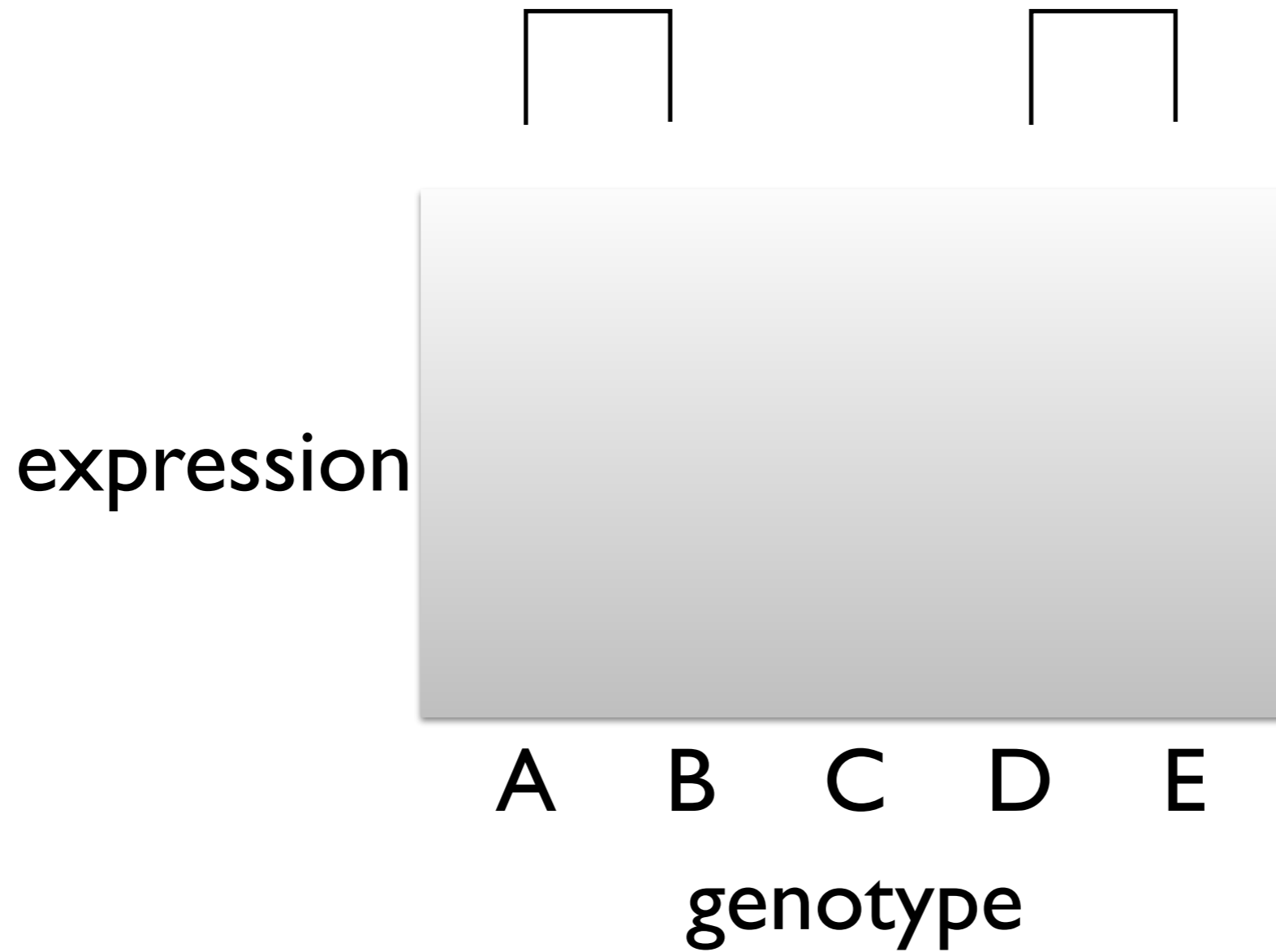
Error bars
reflect 95% CI
(SE or SD would
be appropriate
too)

“The means were significantly heterogeneous (one-way anova, $F(4,35)=7.83$, $P=1.3 \times 10^{-4}$)”.

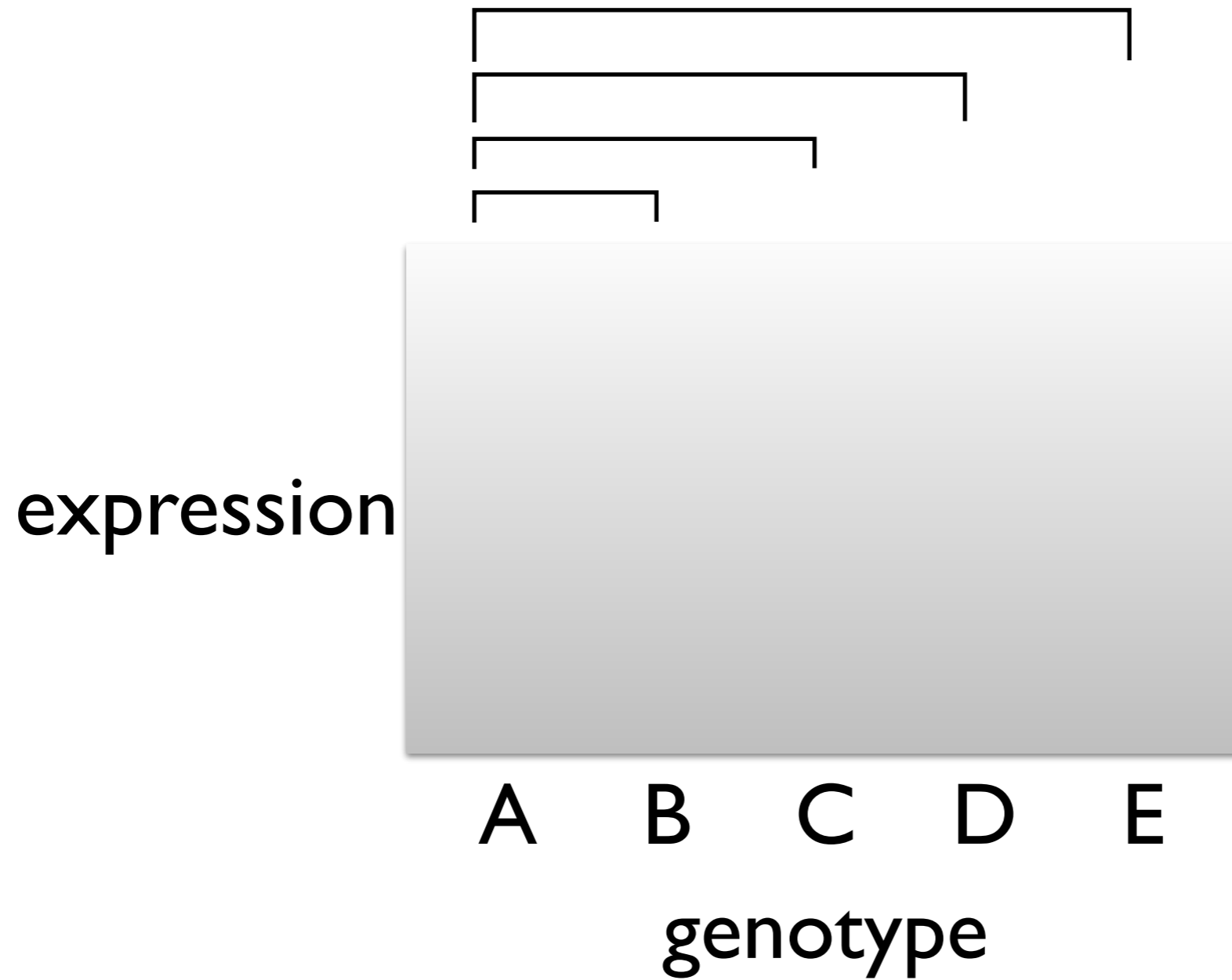
Post tests



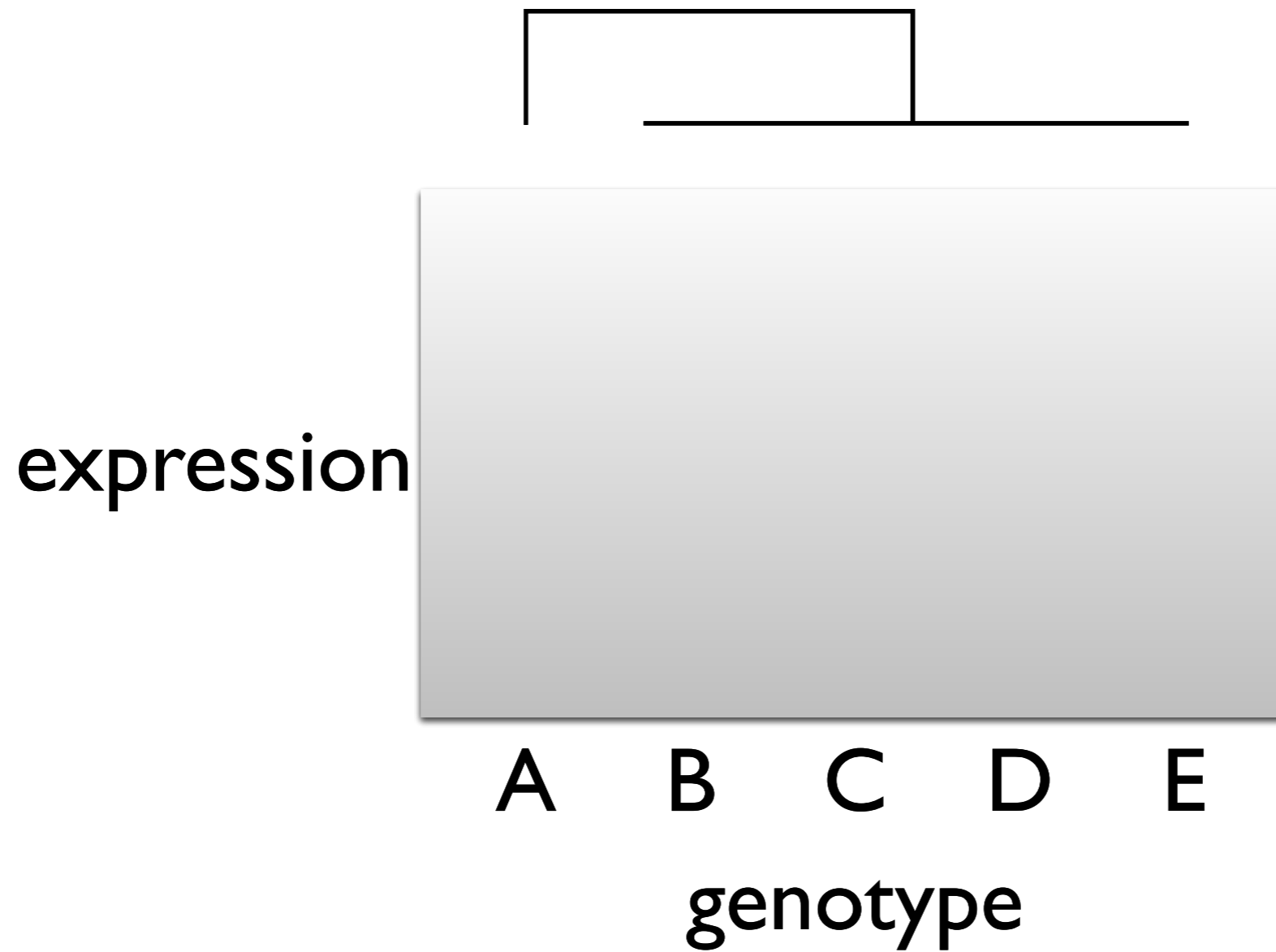
Post tests



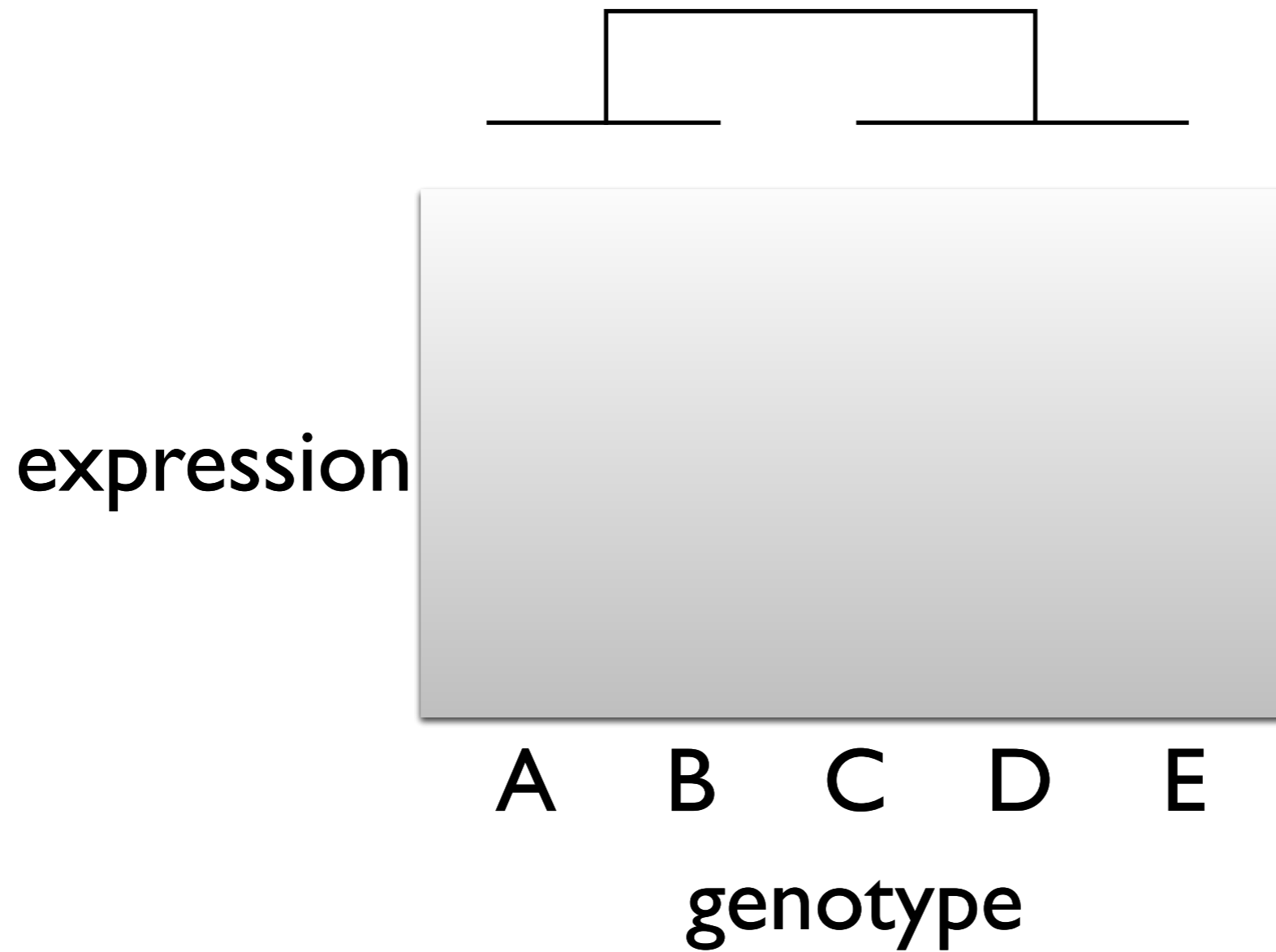
Post tests



Post tests



Post tests



running ANOVA

Analysis of Variance Table

Response: expression

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
genotype	4	10.126	2.53152	7.8323	0.0001283	***
Residuals	35	11.313	0.32322			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Post tests

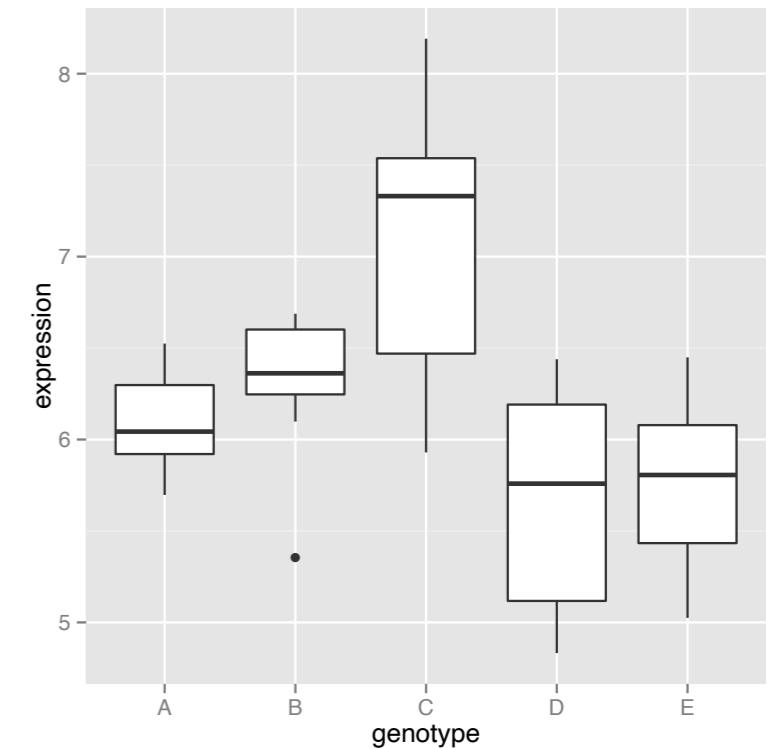
- compare all pairs: Bonferroni, **Tukey**, Student-Newman-Keuls, preferred method depends on number of groups
- **Dunnett**: compares a set of treatments against a single control mean
- all possibilities (contrasts): **Scheffé** test (low power)
- groups naturally ordered: test for trends

all pairs: Tukey

Tukey multiple comparisons of means
95% family-wise confidence level

```
Fit: aov(formula = expression ~ genotype, data = mm)
```

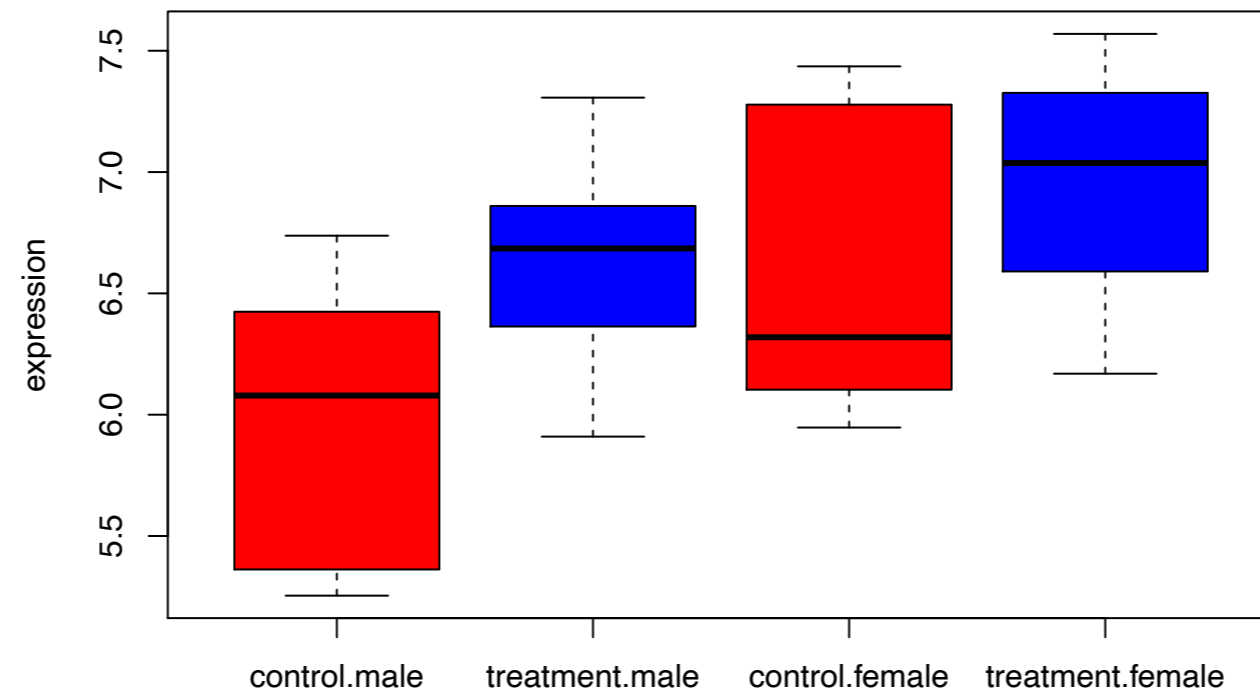
		\$genotype			
	diff	lwr	upr	p adj	
B-A	0.2118143	-0.60545281	1.0290815	0.9441907	
C-A	0.9824239	0.16515675	1.7996910	0.0118801	
D-A	-0.4203216	-1.23758873	0.3969455	0.5826797	
E-A	-0.3413169	-1.15858403	0.4759503	0.7509171	
C-B	0.7706096	-0.04665757	1.5878767	0.0724970	
D-B	-0.6321359	-1.44940305	0.1851312	0.1948625	
E-B	-0.5531312	-1.37039835	0.2641359	0.3131110	
D-C	-1.4027455	-2.22001262	-0.5854783	0.0001806	
E-C	-1.3237408	-2.14100792	-0.5064736	0.0004106	
E-D	0.0790047	-0.73826243	0.8962718	0.9986269	



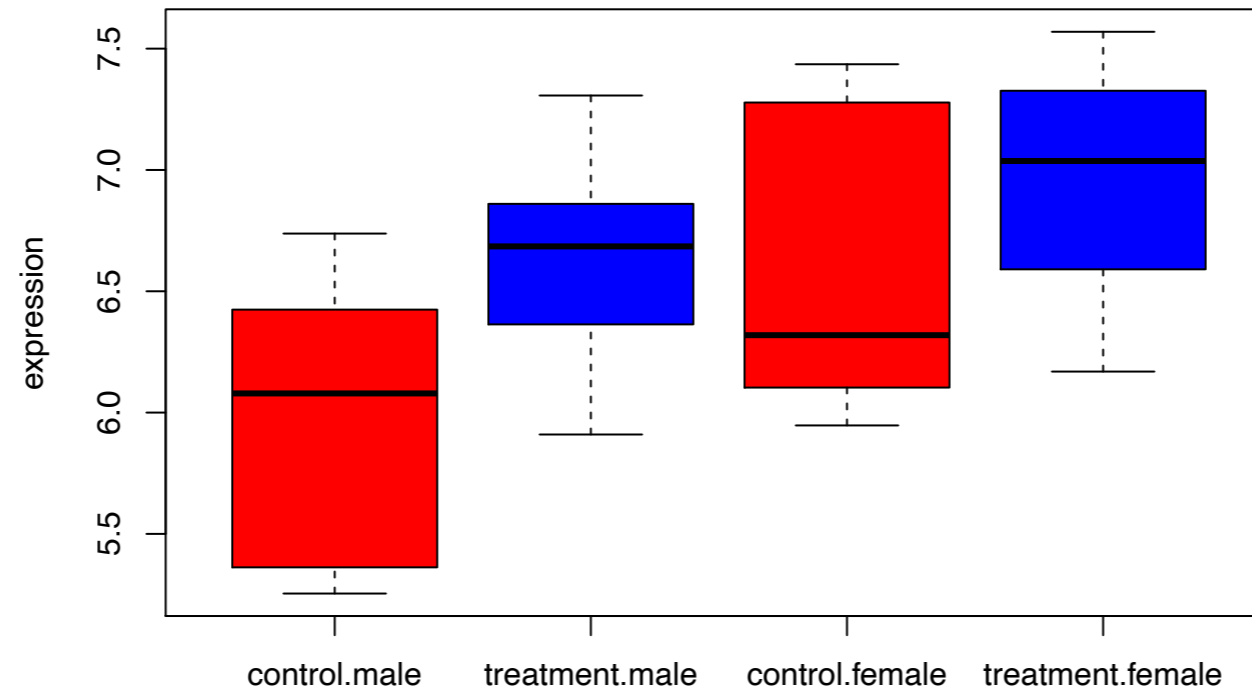
Variations of ANOVA

- non-parametric version of ANOVA: Kruskal Wallis Test
- matched measurements across groups: Repeated-Measures ANOVA

2-way ANOVA



- First Factor differences of means
- Second Factor differences of means
- Interaction of Factor I and Factor II



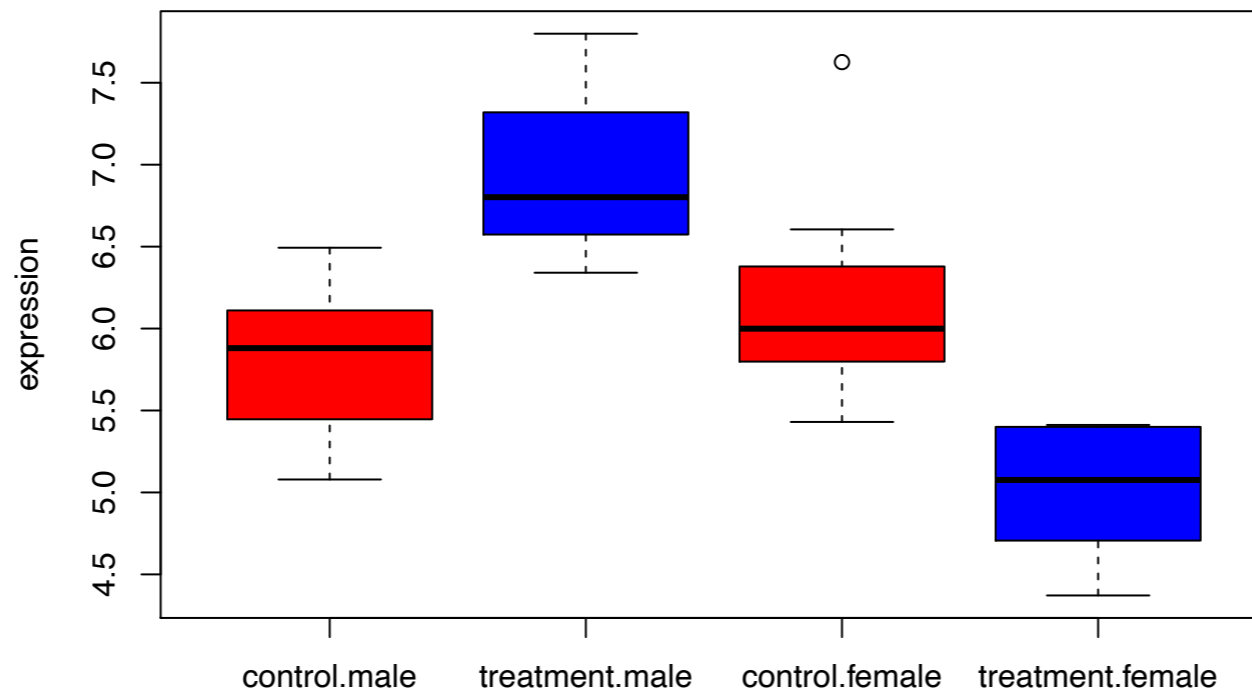
Analysis of Variance Table

Response: expression

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
treatment	1	2.0882	2.08819	7.2159	0.01201	*
gender	1	1.8393	1.83932	6.3560	0.01767	*
treatment:gender	1	0.1873	0.18728	0.6472	0.42791	
Residuals	28	8.1028	0.28939			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

interaction



Analysis of Variance Table

```
Response: expression
      Df  Sum Sq Mean Sq F value    Pr(>F)
treatment  1  0.0010  0.0010  0.0037 0.9521800
gender      1  4.8191  4.8191 17.7482 0.0002369 ***
treatment:gender 1 10.5151 10.5151 38.7257 1.006e-06 ***
Residuals    28  7.6028  0.2715
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

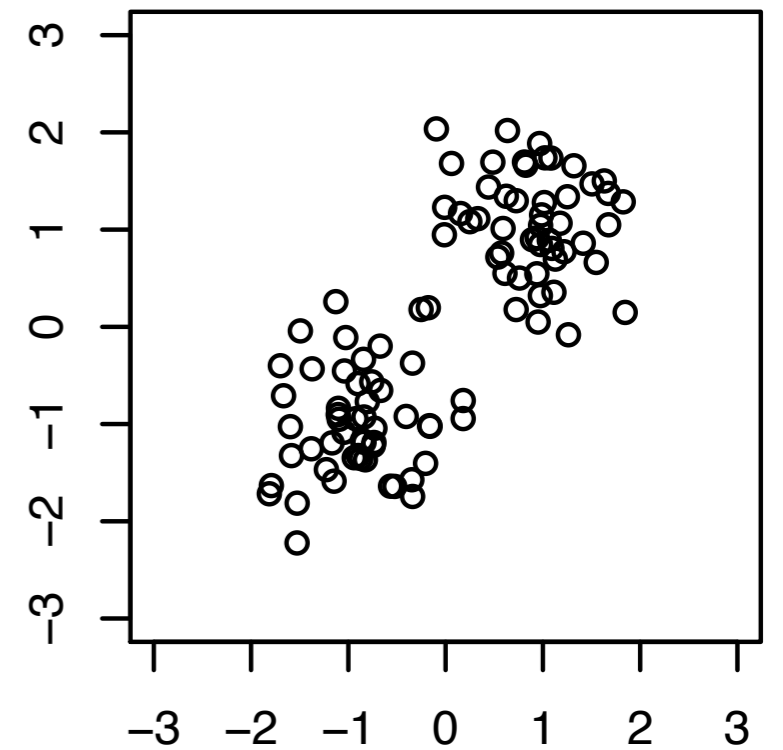
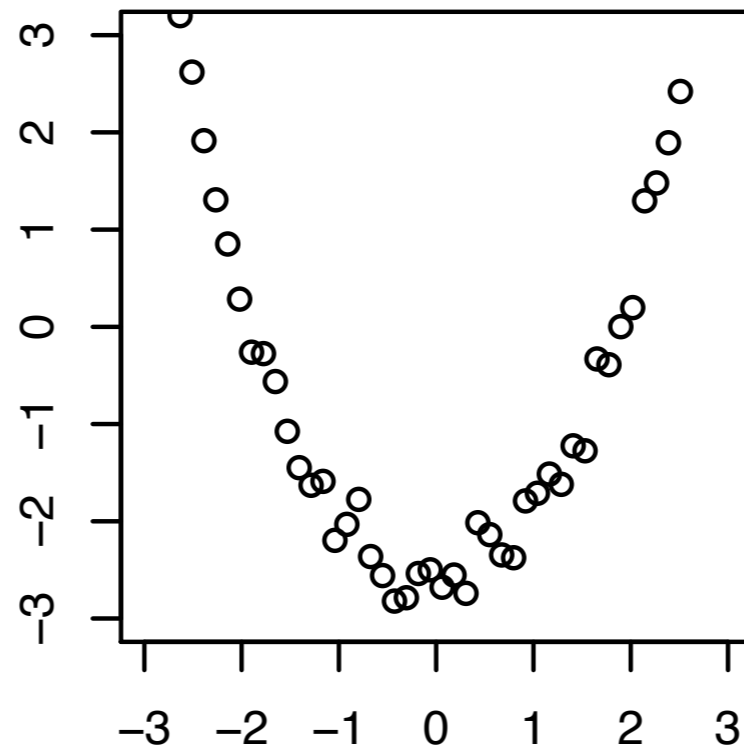
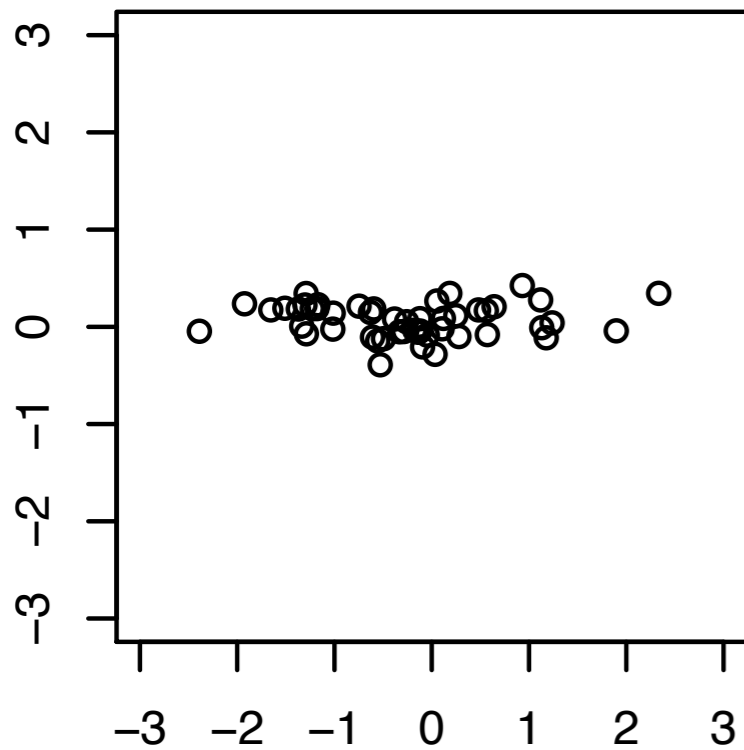
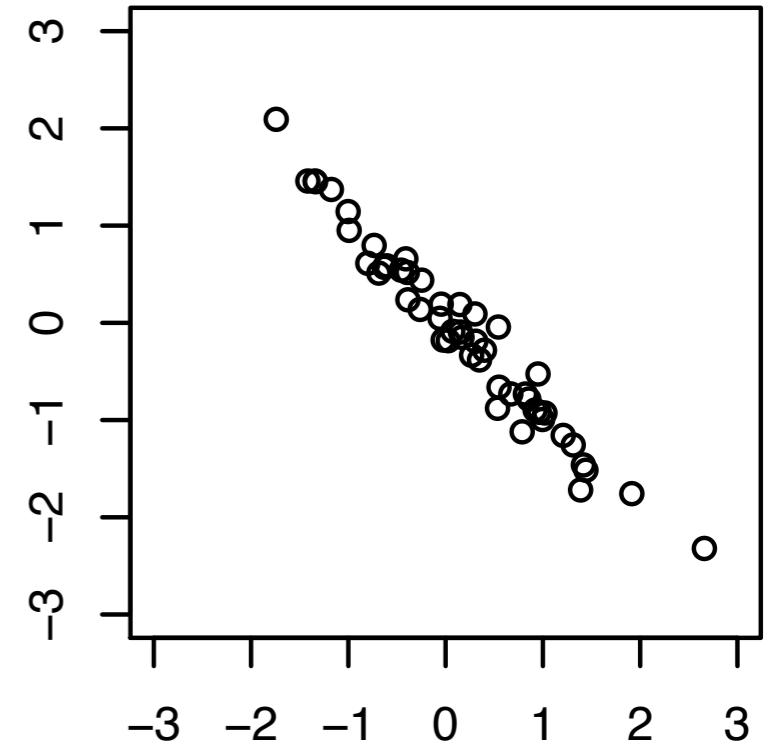
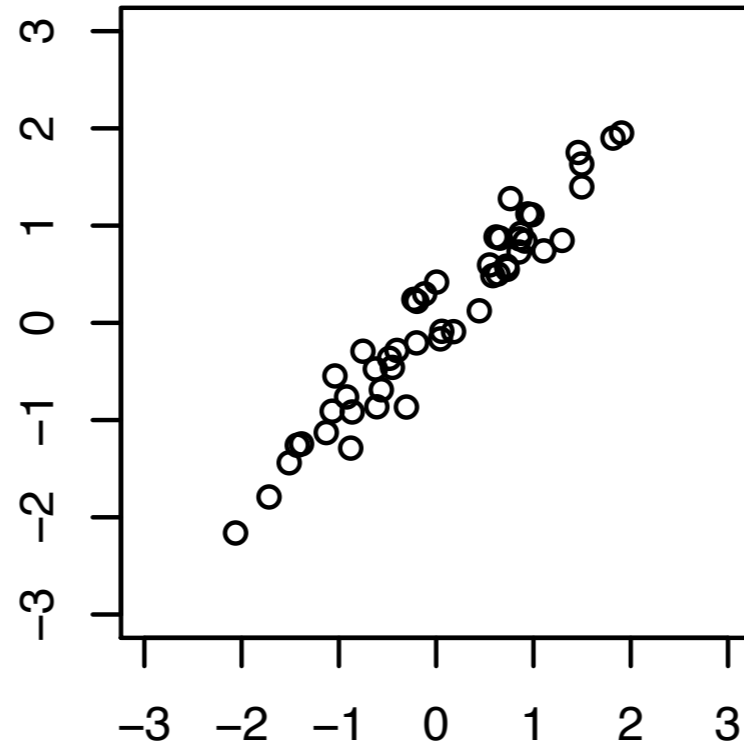
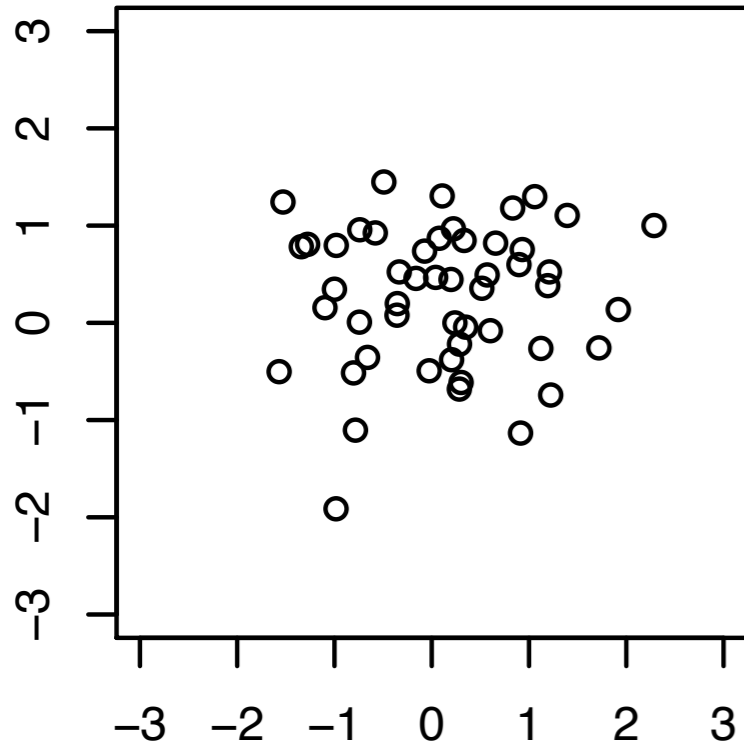
Bivariate Analysis

	x	y
[1,]	0.3019900	-0.6134757
[2,]	0.6567339	0.8198604
[3,]	-0.3538068	0.1979478
[4,]	-1.0974897	0.1558479
[5,]	-0.9836460	-1.9128283
[6,]	0.2854093	-0.2189882

...

Relation of two Variables

Correlations

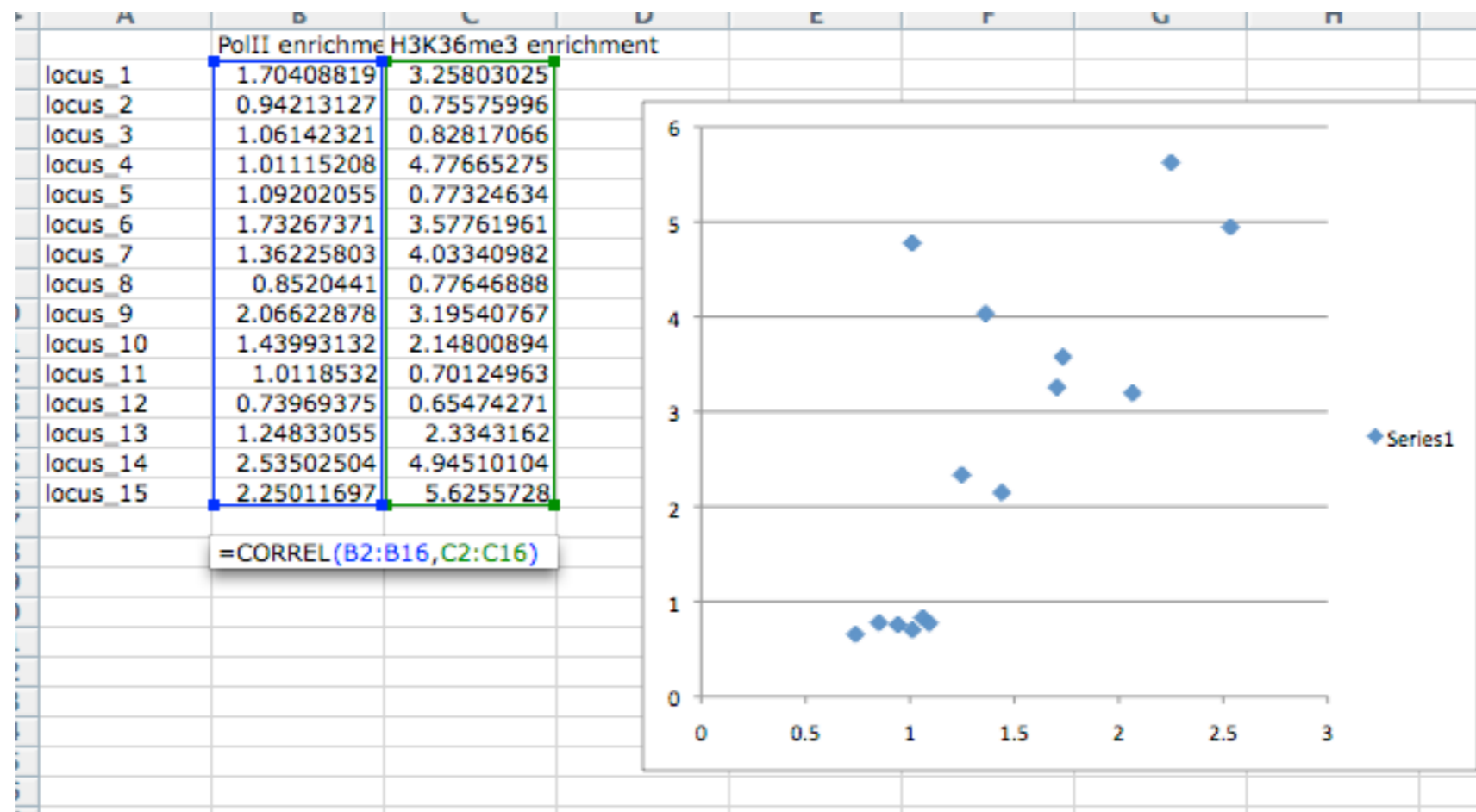


how to **quantify** the relation between 2 continuous variables?

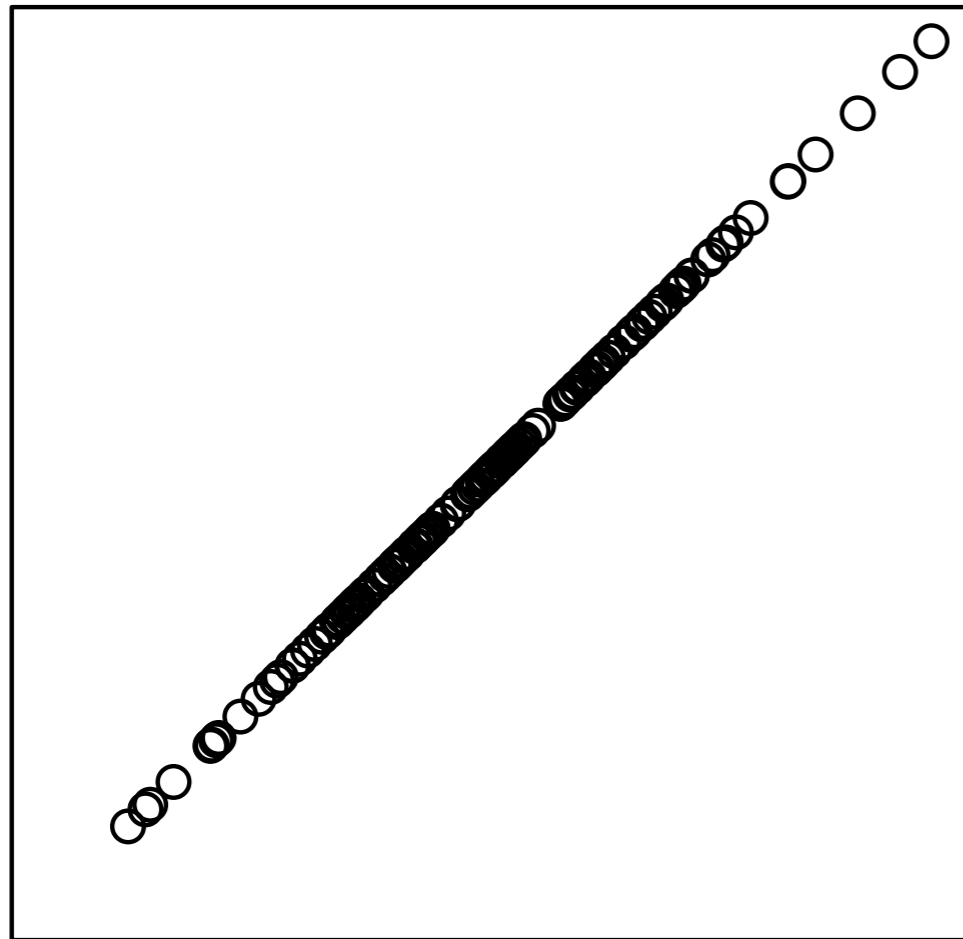
Pearson's Correlation Coefficient

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

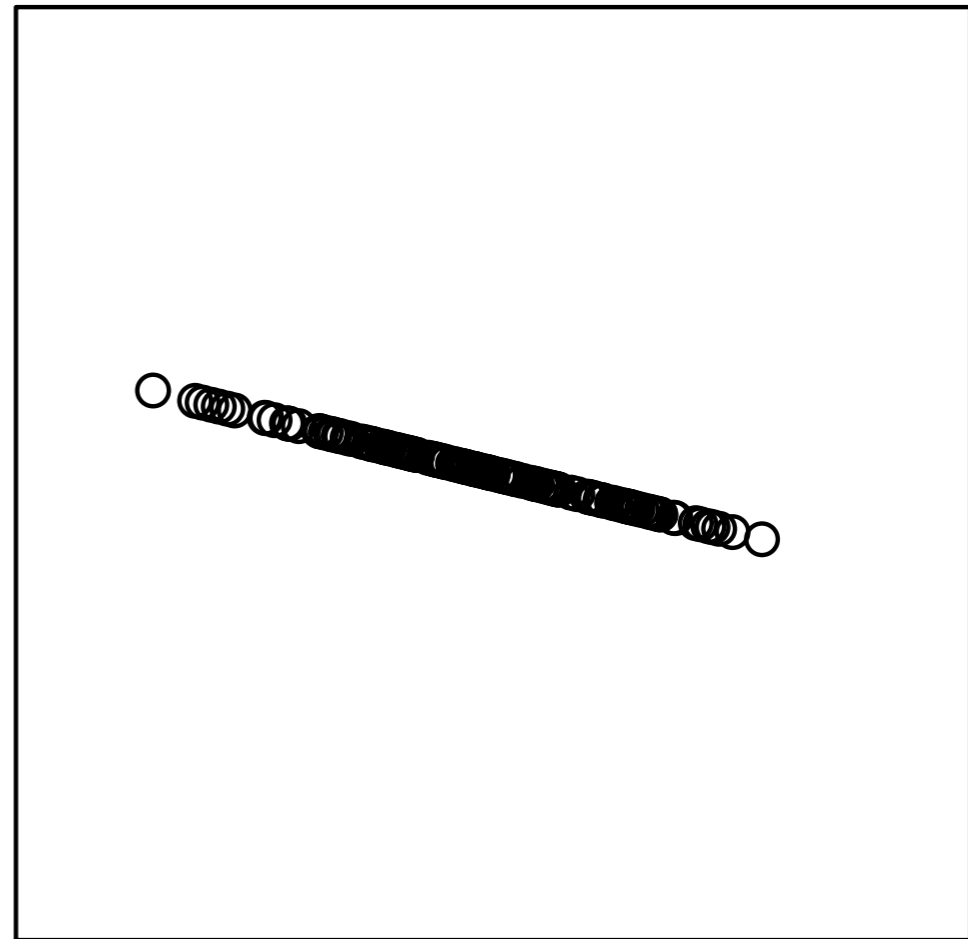
calculation of Pearson correlation in EXCEL



Pearson's Correlation Coefficient

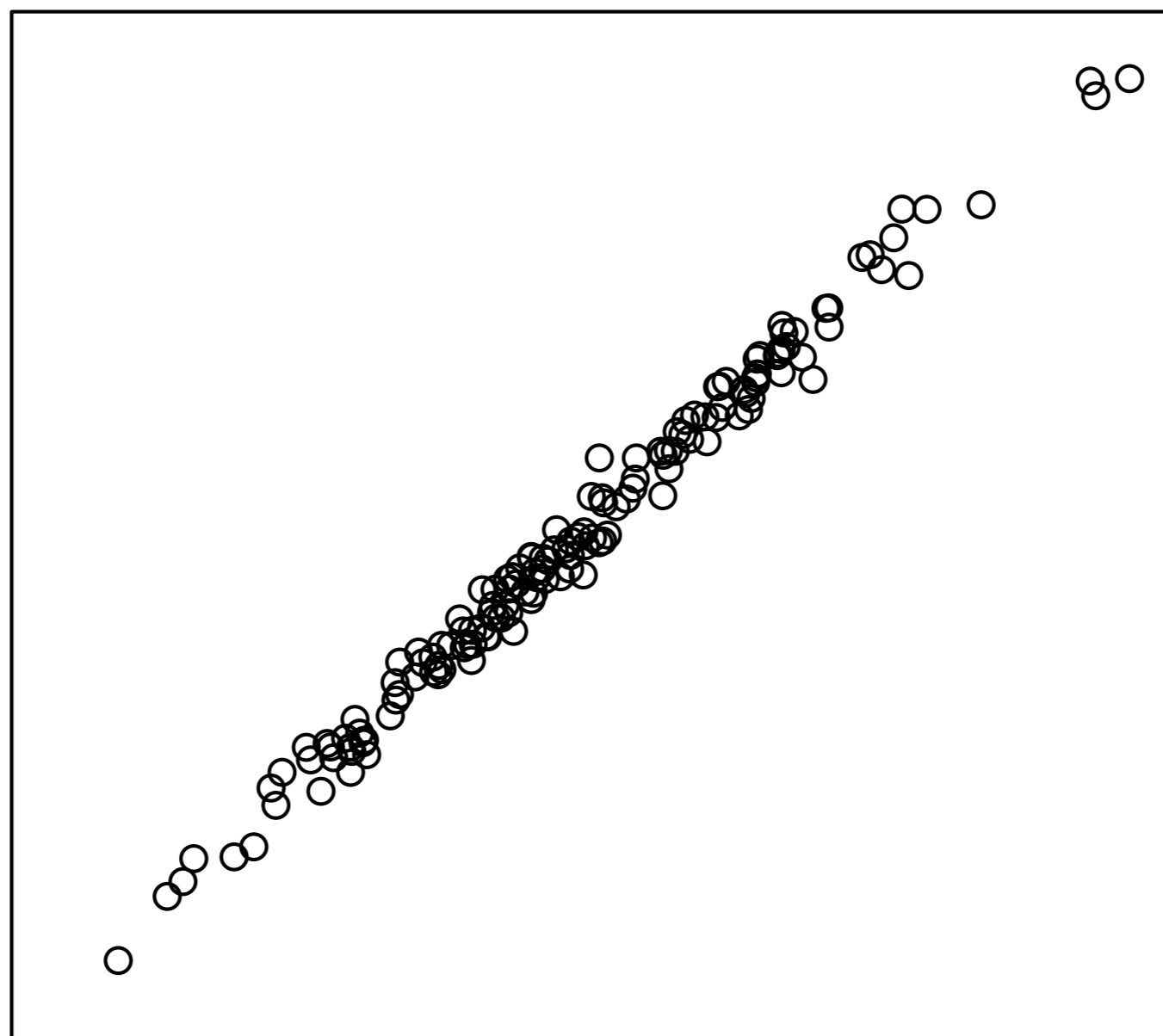


$r = 1$



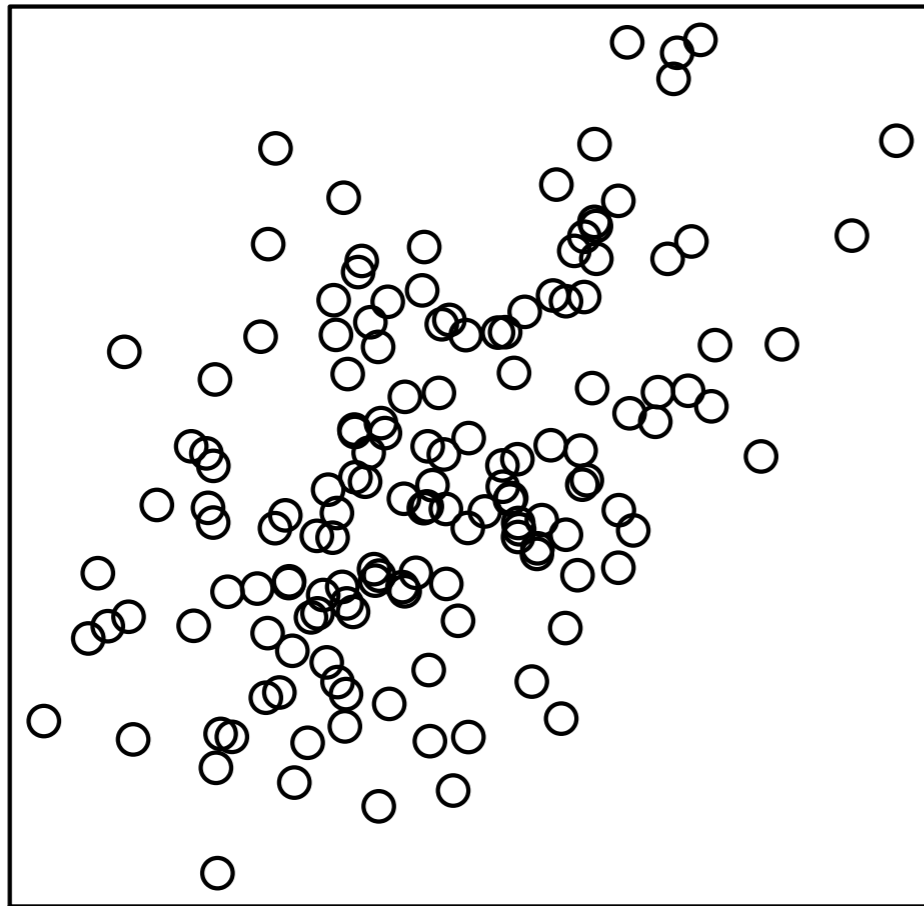
$r = -1$

Pearson's Correlation Coefficient

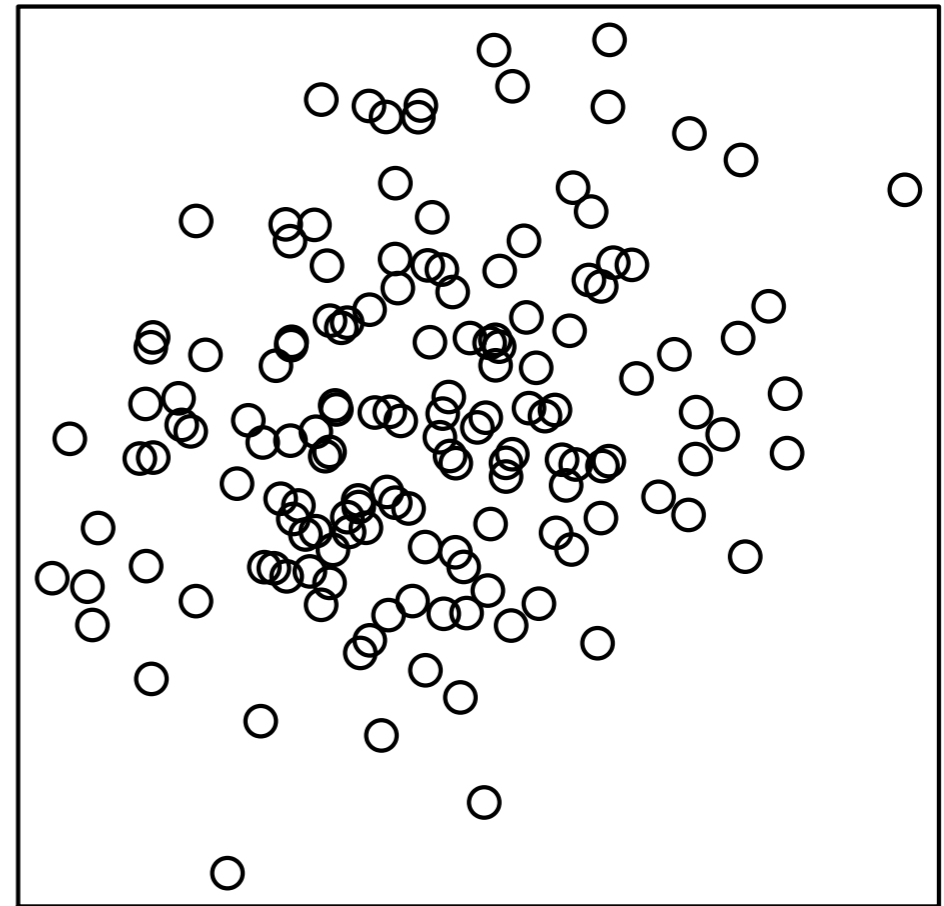


$r = 0.99$

Pearson's Correlation Coefficient

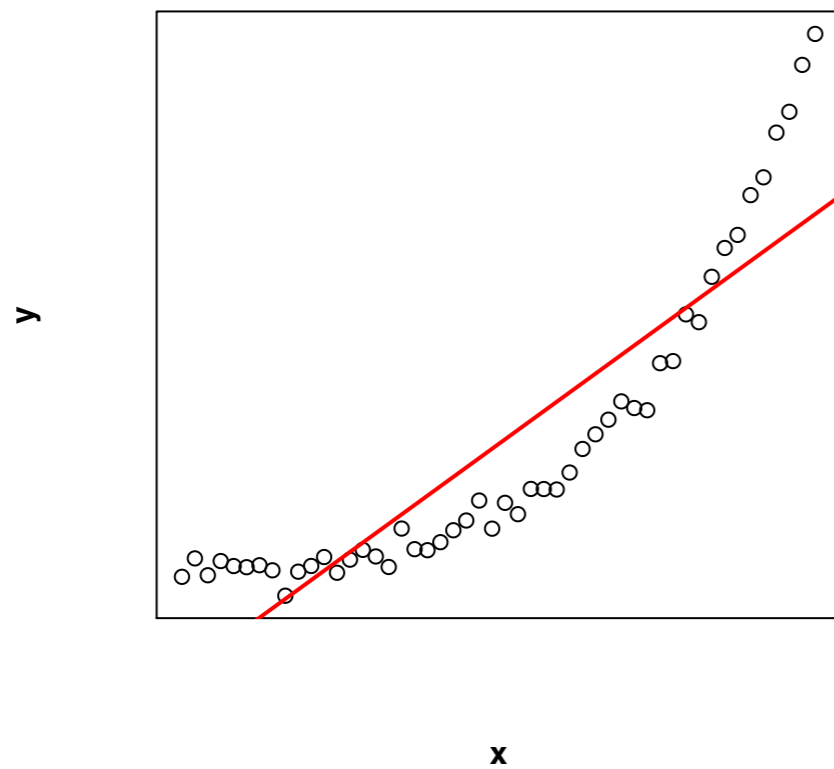


$r = 0.49$

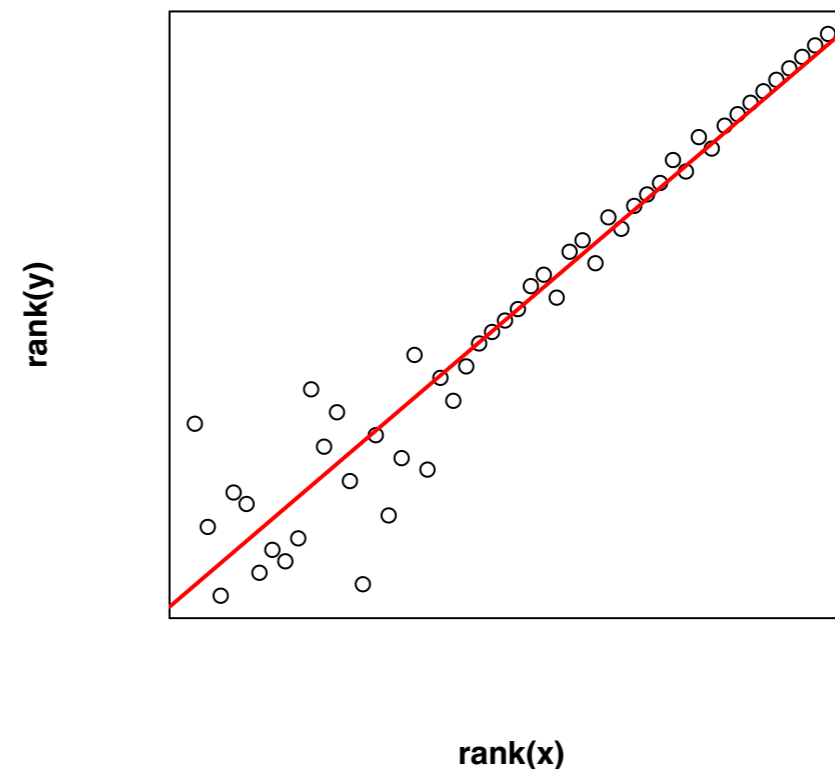


$r = 0.24$

non-linear relationships

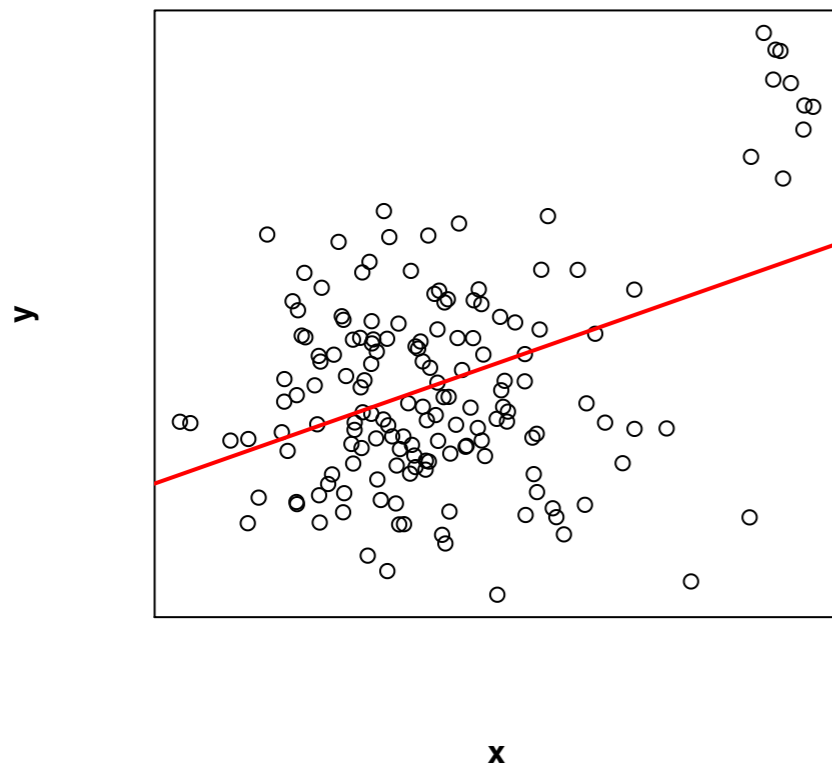


Pearson correlation
 $r=0.88$

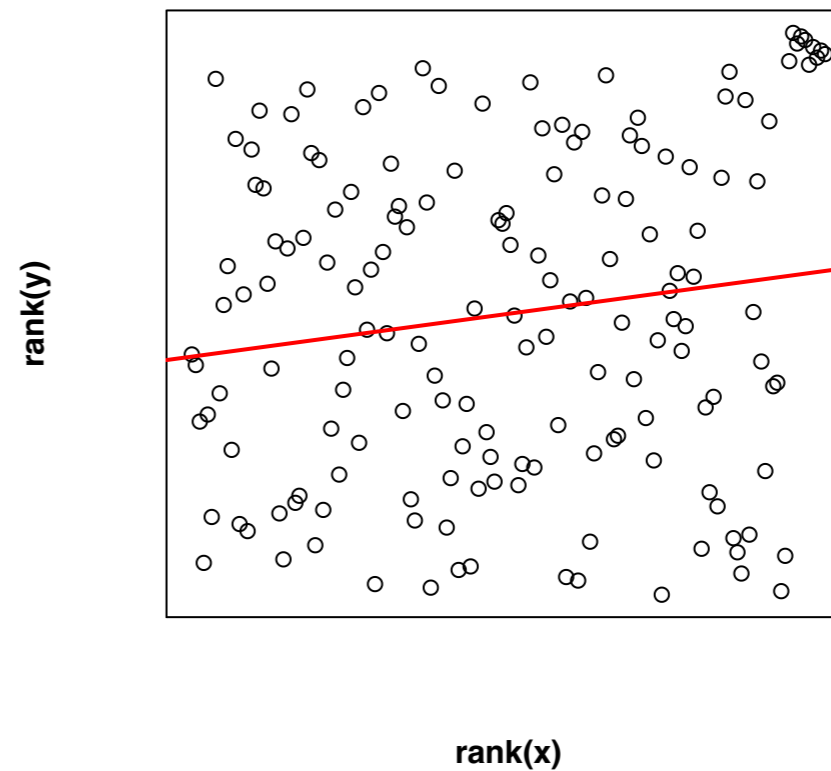


Spearman correlation
 $r_s=0.99$

non-linear relationships



Pearson correlation
 $r=0.42$



Spearman correlation
 $r_s=0.15$

Pearson/Spearman Summary

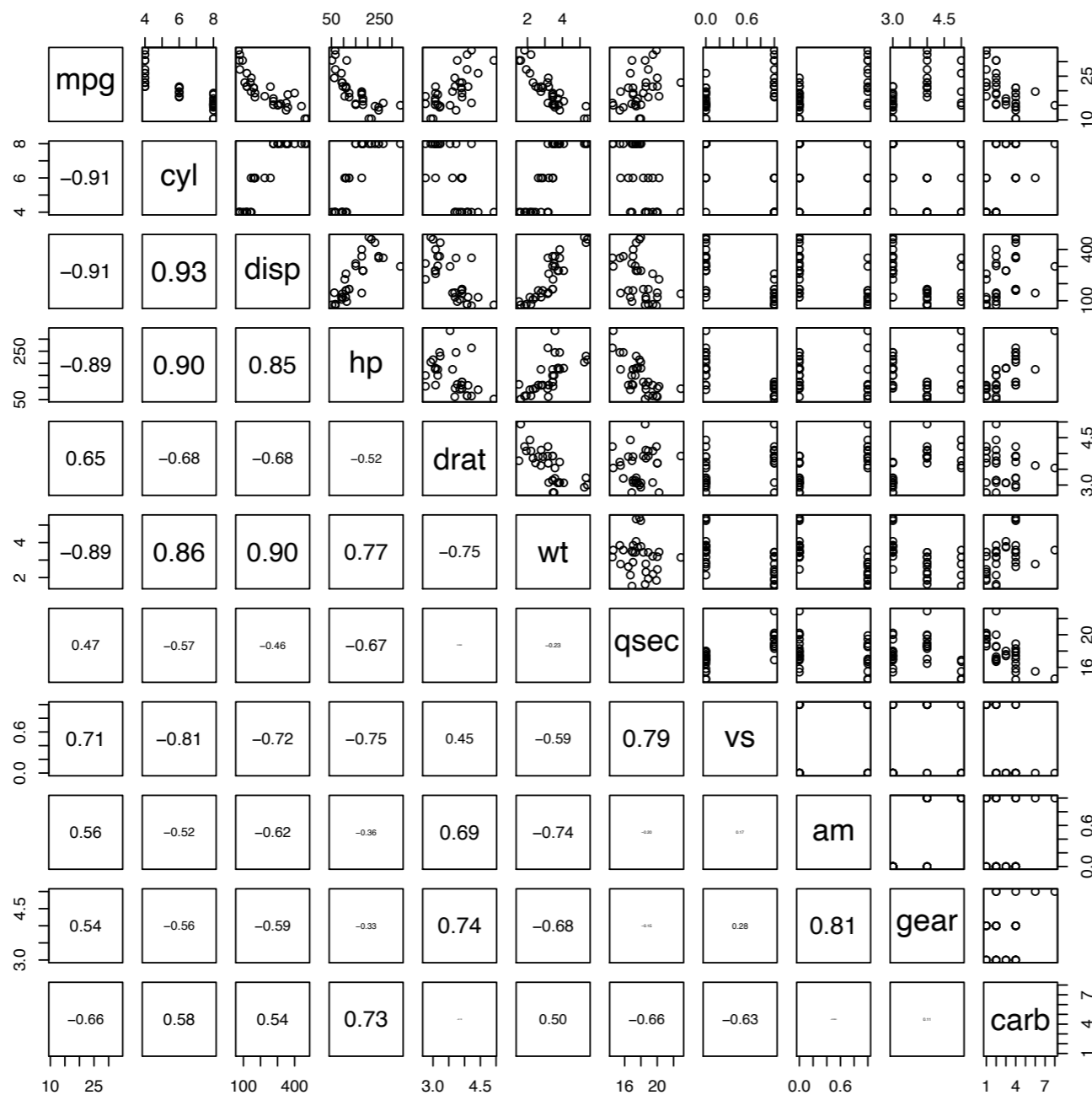
- Pearson correlation is a measure for *linear* dependence
- Spearman correlation is a measure for *monotone* dependence
- The Spearman correlation is *less sensitive* than the Pearson correlation to strong outliers that are in the tails of both samples.
- Correlation coefficients do *not* tell anything about the (non-)existence of a functional dependence.
- Correlation coefficients tell *nothing* about causal relations of two variables X and Y (on the contrary, they are symmetric in X and Y)
- Correlation coefficients hardly tell anything about the shape of a scatterplot

Significance of correlations

- Correlation coefficients are a measure for the strength of a relationship between 2 variables
- That does not tell us anything about the significance of a relationship
- The significance of a correlation is expressed in probability levels (p-values) telling how likely a given correlation coefficient will occur given no relationship in the population.
- Can be calculated easily in R using “cor.test”

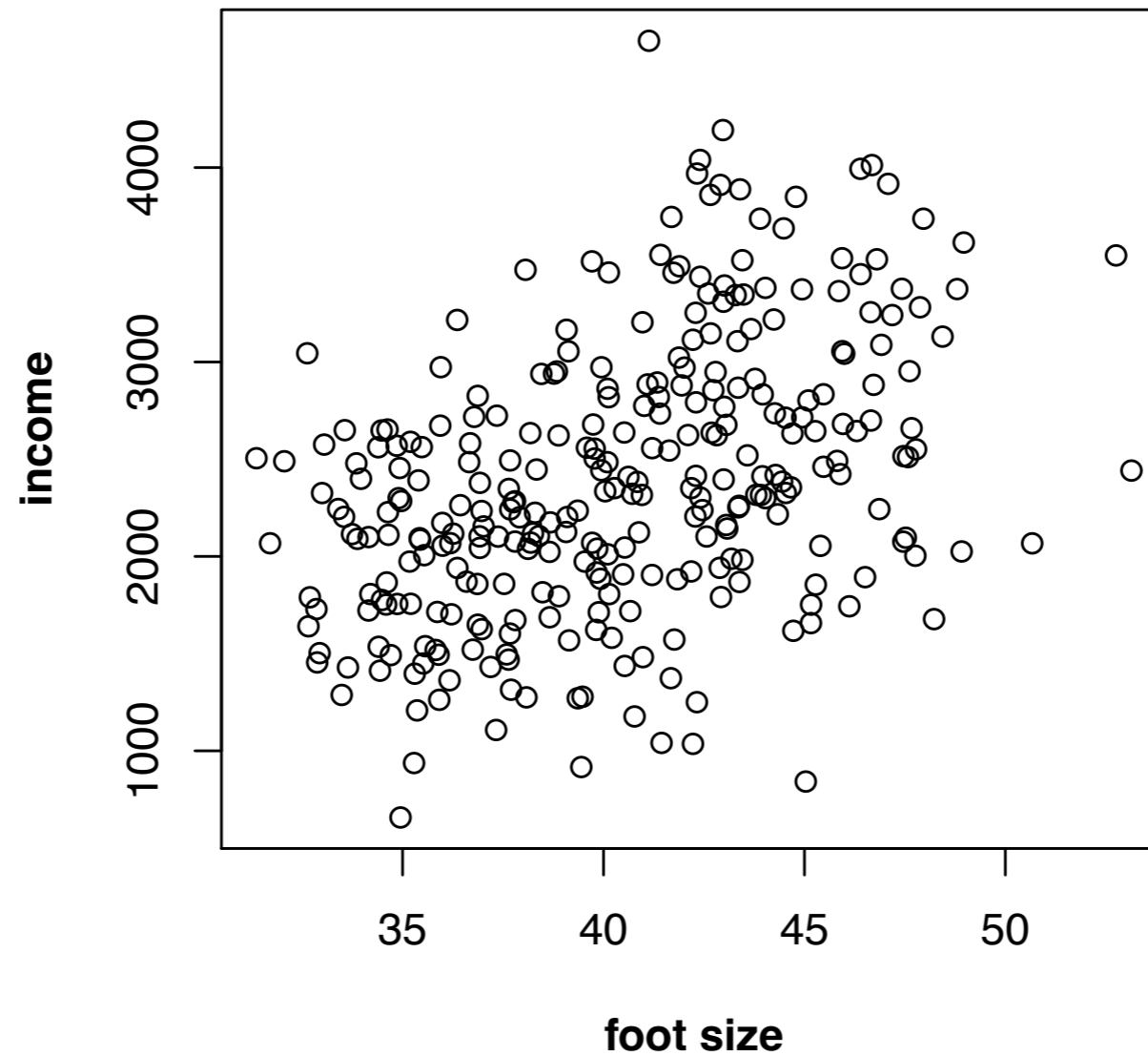
Explorative data analysis using correlations

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1



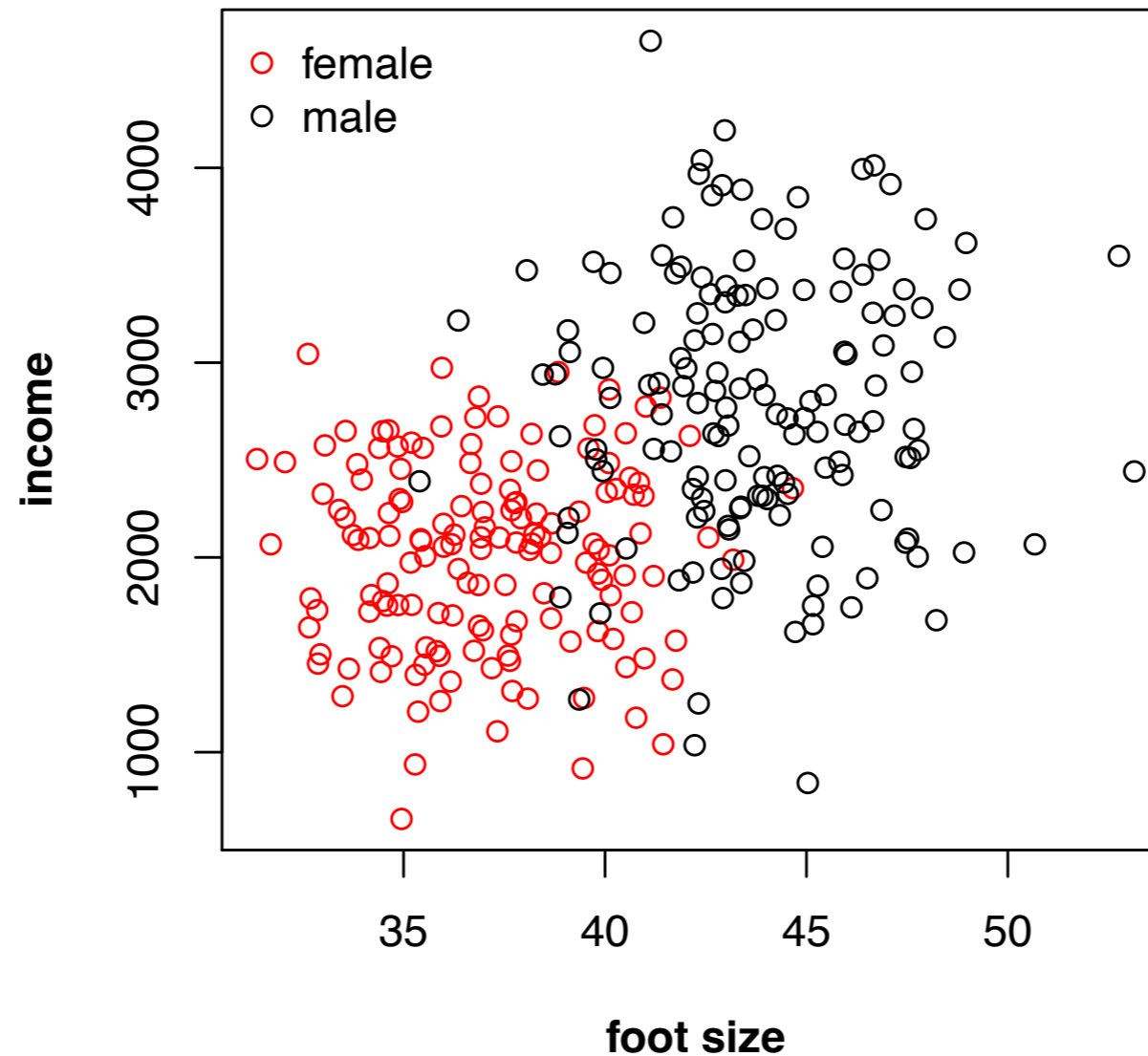
mpg – Miles/(US) gallon
 cyl – Number of cylinders
 disp – Displacement (cu.in.)
 hp – Gross horsepower
 drat – Rear axle ratio
 wt – Weight (lb/1000)
 qsec – 1/4 mile time
 vs – V/S
 am – Transmission (0 = automatic, 1 = manual)
 gear – Number of forward gears
 carb – Number of carburetors

Confounding - watch out!



Pearson correlation
 $r=0.42$

Confounding - watch out!



Confounding: A variable that „explains“ (part of) the dependence of two others

responsible research

statistics don't lie but liars use statistics

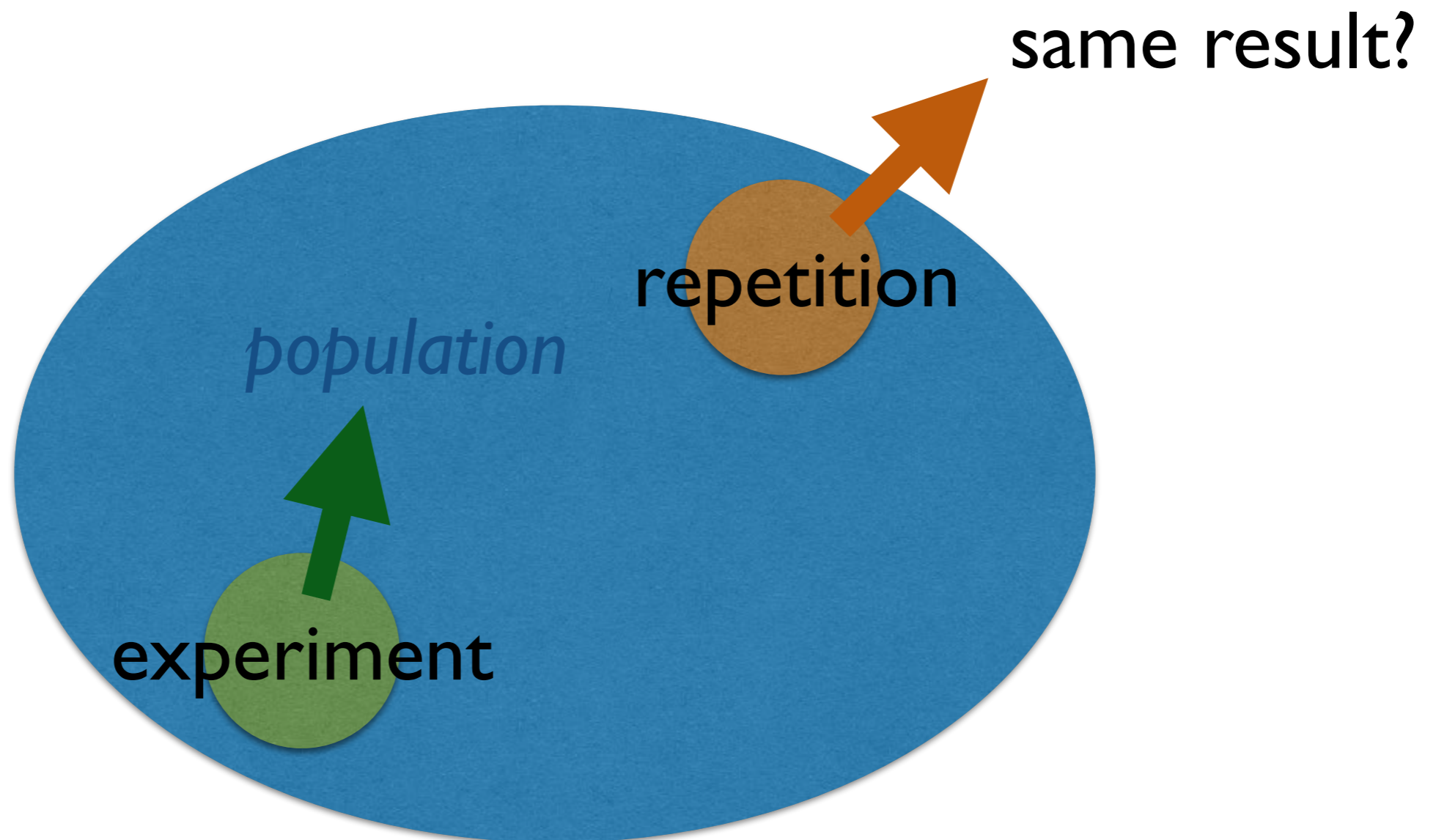


© Scott Adams, Inc./Dist. by UFS, Inc.

most (90-95%) of the published pre-clinical research findings are wrong (irreproducible)

- Ioannidis JPA. 2005. Why most published research findings are false. PLoS Med 2: e124.
- Begley CG, Ellis LM. 2012. Drug development: Raise standards for preclinical cancer research. Nature 483: 531–533.
- irreproducibility correlates with:
 - inappropriate application of statistical procedures
 - low statistical power
 - inappropriate experimental design
 - ...

Estimating reproducibility



Replicability



Reproducibility

Reproduction of the original results using the same protocol/reagents/tools

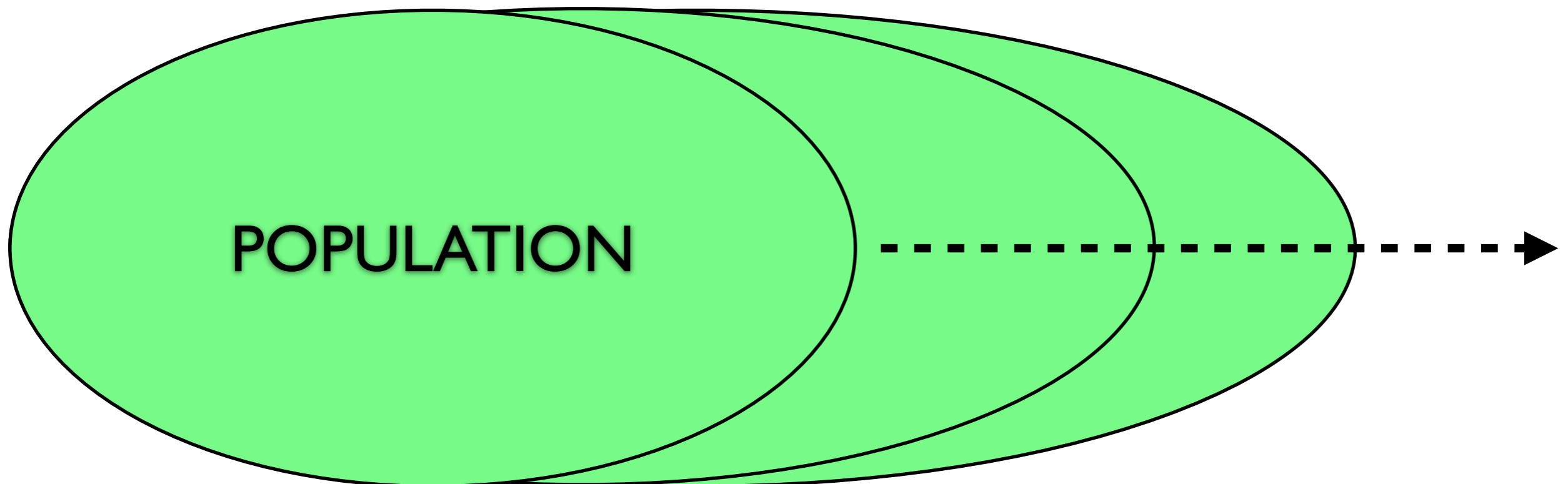
by the same person

by a different person in the lab

by a different person outside the lab

Reproduction using different reagents/tools but the same protocol by a different person outside the lab

Reproduction just based on text description



How to avoid sampling bias?

- *blinding*: the person conducting the experiment should e.g. not be aware of whether control or treatment is applied
- *randomisation*: the samples should be assigned randomly to experimental groups
- *exclusion* criteria should be defined if exclusion of data is likely to happen.
- *confounding* factors have to be identified and controlled for

A QPR show case

The Journal of Neuroscience, April 16, 2014 • 34(16):5529–5538 • 5529

Neurobiology of Disease

Cannabis Use Is Quantitatively Associated with Nucleus Accumbens and Amygdala Abnormalities in Young Adult Recreational Users

The Washington Post

Morning Mix

Even casually smoking marijuana can change your brain, study says

confounding

Table 1. Participant demographics

	CON (<i>n</i> = 20)	MJ (<i>n</i> = 20)	<i>p</i> -value
Sex (M/F)	9 M/11 F	9 M/11 F	N/A
Age	20.7 (1.9)	21.3 (1.9)	0.30
Years of education	14.3 (3.4)	12.6 (4.8)	0.20
STAI ^a			
State	28.9 (7.94)	27.7 (7.38)	0.65
Trait	29.8 (7.32)	29.5 (5.56)	0.89
HAM-D ^b	0.80 (1.40) [range: 0–5]	1.10 (1.37) [range: 0–5]	0.50
TIP ^c			
Extroversion	10.9 (2.36)	10.7 (2.13)	0.78
Agreeableness	10.8 (2.47)	10.7 (1.81)	0.94
Conscientiousness	11.9 (2.08)	11.7 (2.13)	0.76
Emotional stability	10.5 (2.52)	11.4 (2.64)	0.27
Openness	12.1 (1.90)	12.4 (1.61)	0.57
Substance use			
Alcohol			
No. alcoholic drinks/week	2.64 (2.38)	5.09 (4.69)	0.10
AUDIT score	3.30 (1.78)	5.50 (2.21)	0.05
Cigarettes			
No. of occasional smokers ^d	0	7	N/A
No. of daily smokers	0	1	N/A
Marijuana			
No. days/week	0	3.83 (2.36)	N/A
No. joints/week	0	11.2 (9.61)	N/A
No. joints/occasion	0	1.80 (0.77)	N/A
No. smoking occasions/day	0	1.80 (0.70)	N/A
Age of onset (years)	—	16.6 (2.13)	N/A
Duration of use (years)	—	6.21 (3.43)	N/A

All values are expressed in means and SDs. CON, controls; MJ, marijuana users.

^aState Trait Anxiety Inventory Form (Spielberger et al., 1983).

^bHamilton Depression Rating Scale (Hamilton, 1960).

^cTen-Item Personality Inventory (Gosling et al., 2003).

^dOccasional smokers reported from 1 cigarette/week to 1 cigarette every 3 months.

types of research

EXPLORATORY

- hypothesis generating
- no/little prior information on effects, frequently many endpoints measured (multiple testing)
- often not complying with elementary rules of sampling and experimental layout (e.g. sequential sampling, multiple testing)
- statistical testing will yield highly problematic results (low power, high error rate), potentially irreproducible

CONFIRMATORY

- performed to confirm hypotheses
- solid prior knowledge on effects
- involves prior power analysis, thoughtful experimental layout
- generates more reliable statistical test results, potentially reproducible

Experimental Design

If your experiment needs statistics, you ought to have done a better experiment - Ernest Rutherford

If your statistics should be any valid, you have to plan and perform experiments properly - Anonymous

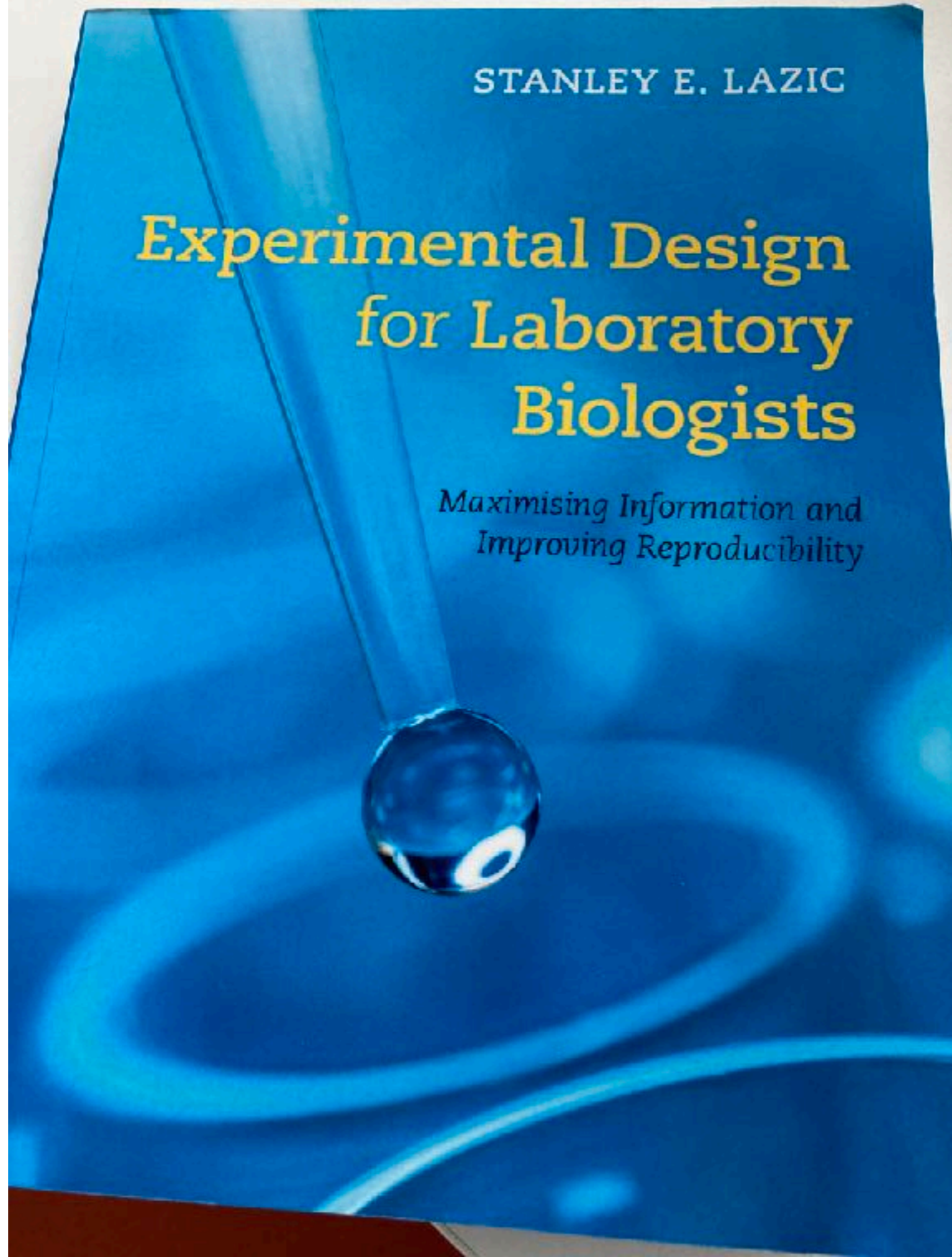
Experimental Design

- Design of experiments, or experimental design, is the design of all information-gathering exercises where variation is present, whether under the full control of the experimenter or not.
- One central aim is to minimize random and systematic error contribution to the variation, such that the fluctuations of the dependent variable (the measurement) are maximally related to the levels of the independent variable (the treatment)
- Valid inferences on the behaviour of an entire population should be derived.

STANLEY E. LAZIC

**Experimental Design
for Laboratory
Biologists**

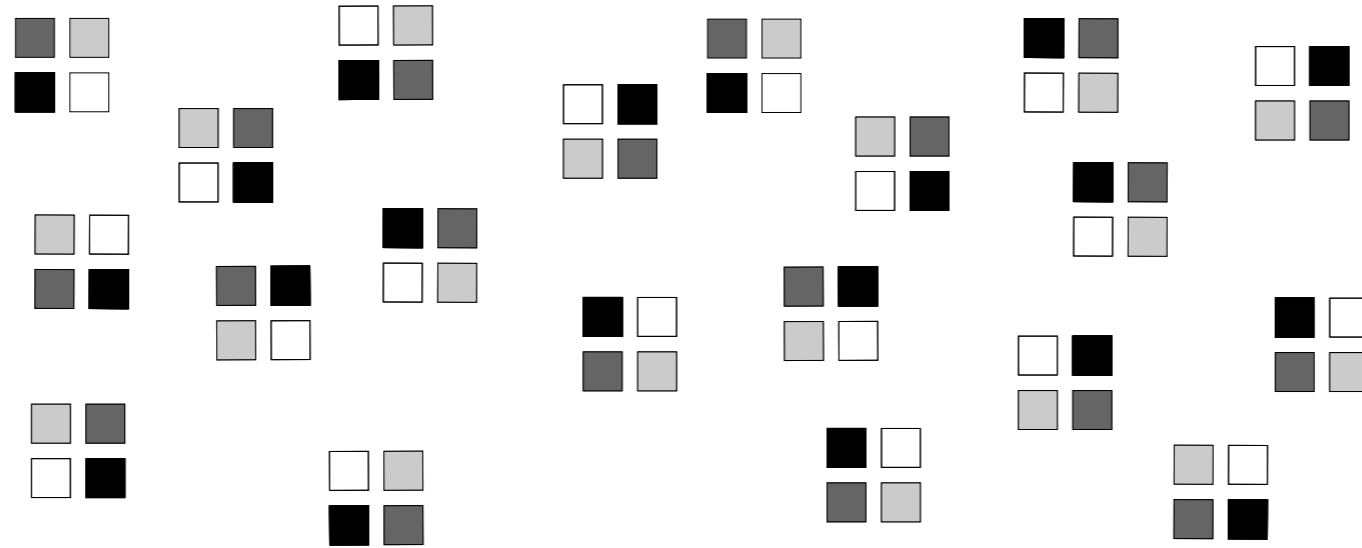
*Maximising Information and
Improving Reproducibility*



experiment flow chart

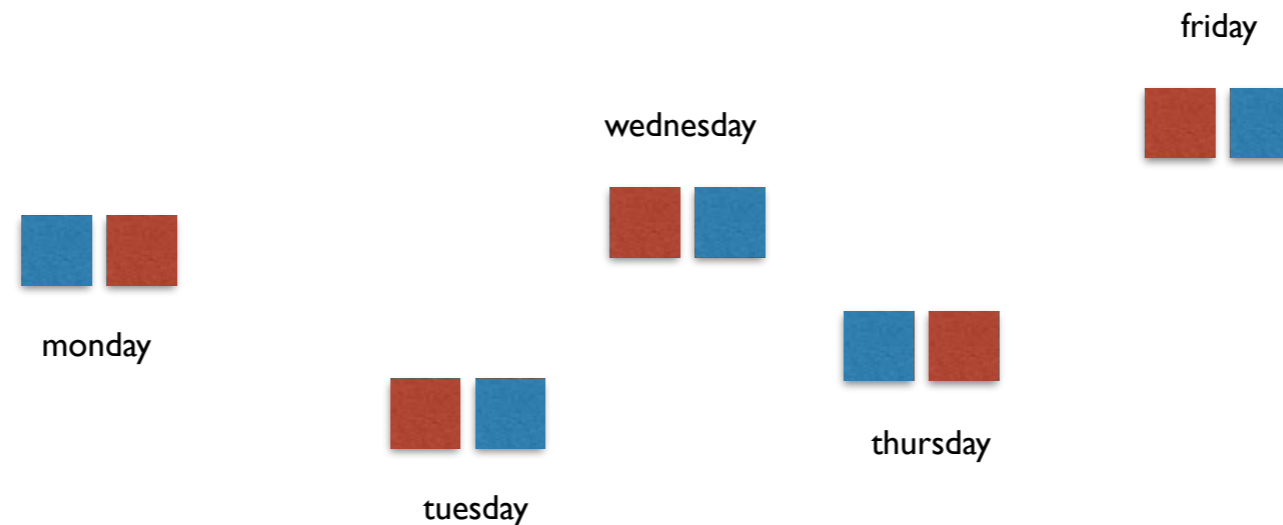
- formulate a hypothesis *before* data collection
- design an experiment to tests this hypothesis
 - ideally this experiment should be a comparative one (2 states)
 - define what you measure (dependent variable), the link between the (proxy) variable and the biological model.
 - make up your mind about the sample size (power analysis) and the statistics you want to apply
 - consider potential sources of error and how you can minimise them
- perform experiment
- analyse your data
- consider to perform a completely different experiment that can confirm your finding

a well designed experiment



- randomised block design
- ANOVA with fixed effect (treatment) and random effect (block)
- Problem: randomisation and statistical testing should involve an experienced statistician

the ideal design



- randomised block design, only 2 factor levels (control, treatment)
- suited to control for day-to-day fluctuations which are very common. Ideally one would change reagents, batches of cells etc. between the blocks as well. Every block a new batch, every block new reagents.
- paired t-test

N is (too) small, what can you do?

- Improve experimental design
 - simple comparative studies (2-group) have higher power than complex studies
 - reduce systematic errors by e.g. random block design
- Improve the power of statistical test
 - paired tests instead of unpaired tests (requires appropriate experimental design)
 - avoid making comparisons that are of no interest