Reproducible Research

Dr. Tobias Straub
BMC Bioinformatics, LMU, Munich

Over the past decade, before pursuing a particular line of research, scientists (including C.G.B.) in the haematology and oncology department at the biotechnology firm Amgen in Thousand Oaks, California, tried to confirm published findings related to that work. Fifty-three papers were deemed 'landmark' studies (see 'Reproducibility of research findings'). It was

tive clinical uses for existing therapeutics. Nevertheless, scientific findings were confirmed in only 6 (11%) cases. Even knowing the limitations of preclinical research, this

cer research.

In studies for which findings could be reproduced, authors had paid close attention to controls, reagents, investigator bias and describing the complete data set. For results that could not be reproduced, however, data were not routinely analysed by investigators blinded to the experimental versus control groups. Investigators frequently presented the results of one experiment, such as a single Western-blot analysis. They sometimes

Begley CG, Ellis LM. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* **483**: 531–533.

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2

Nevertheless, most new discoveries will continue to stem from hypothesis-generating research with low or very low pre-study odds. We should then acknowledge that statistical significance testing in the report of a single study gives only a partial picture, without knowing how much testing has been done outside the report and in the relevant field at large. Despite a large statistical literature for multiple testing corrections [37], usually it is impossible to decipher how much data dredging by the reporting authors or other research teams has preceded a reported research finding. Even if determining

Ioannidis JPA. 2005. Why most published research findings are false. *PLoS Med* **2**: e124.

Experimenter's contribution to reproducible research

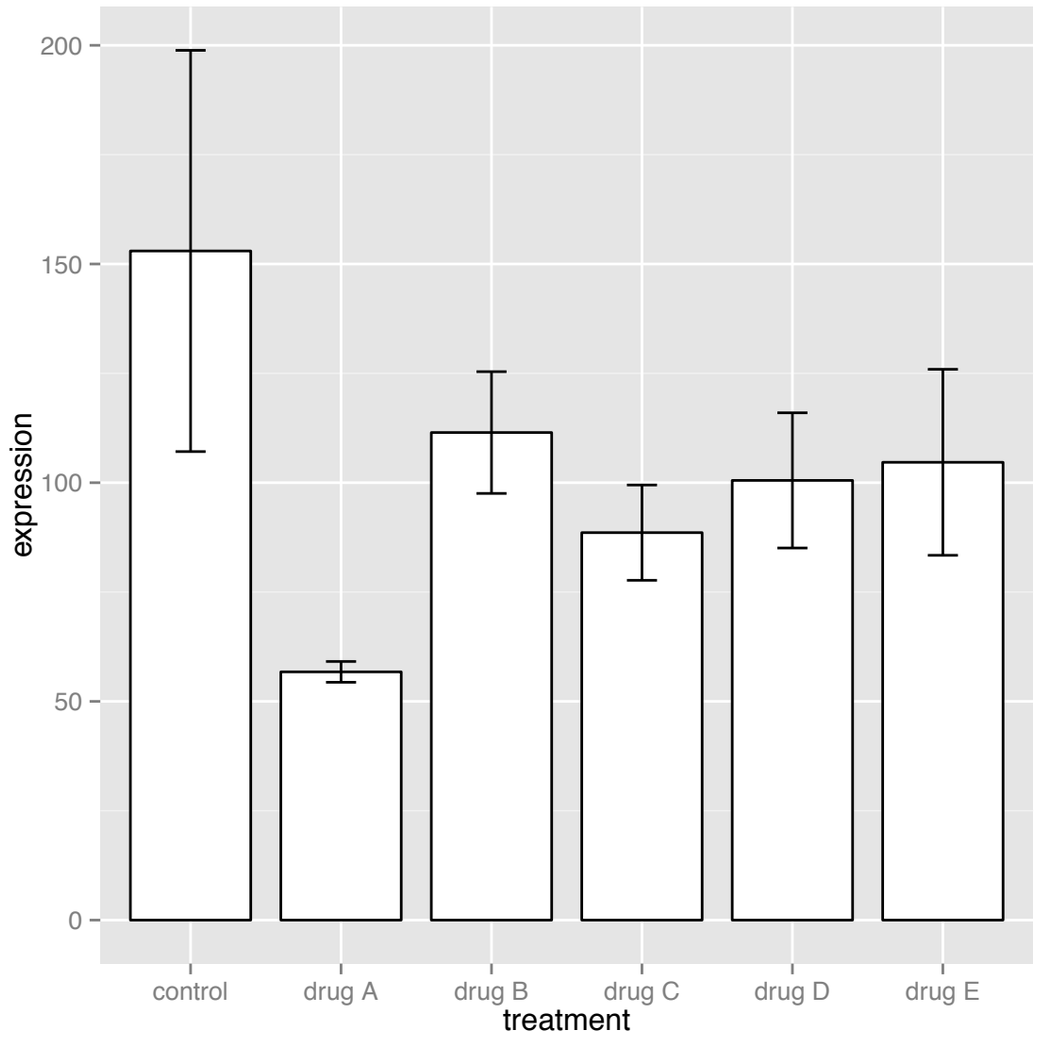
- Experimental Design
- Statistics
- Documentation
- Interpretation

Example

Example

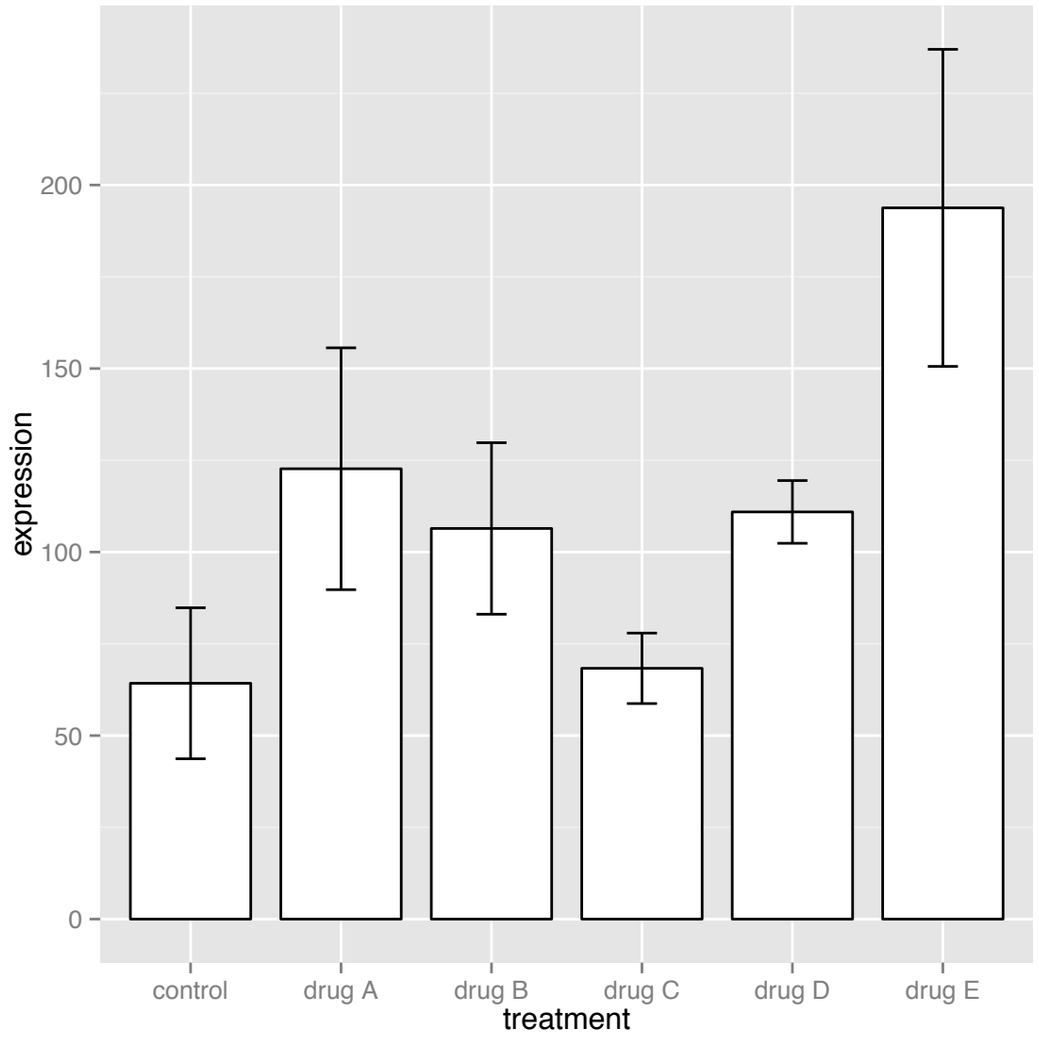
- test a couple of drugs on whether they affect the expression of a gene
- quick shot: qPCR with technical replicates

a first test

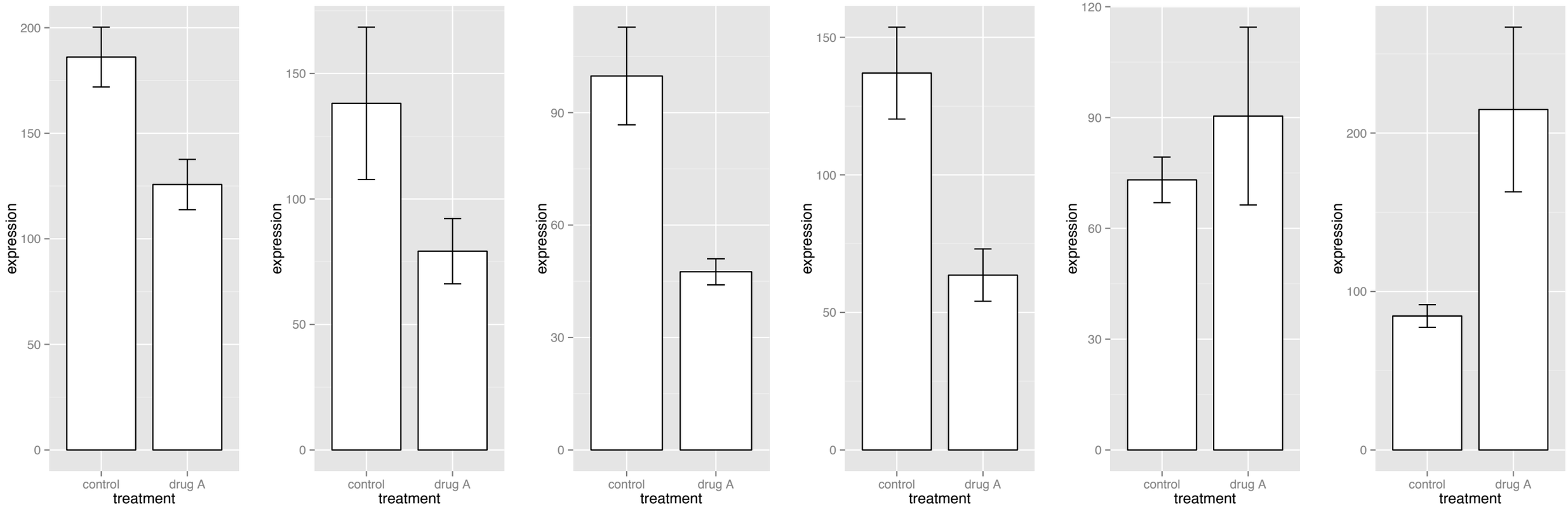


"Wow, drug A shows a significant effect, the error bars do not overlap!"

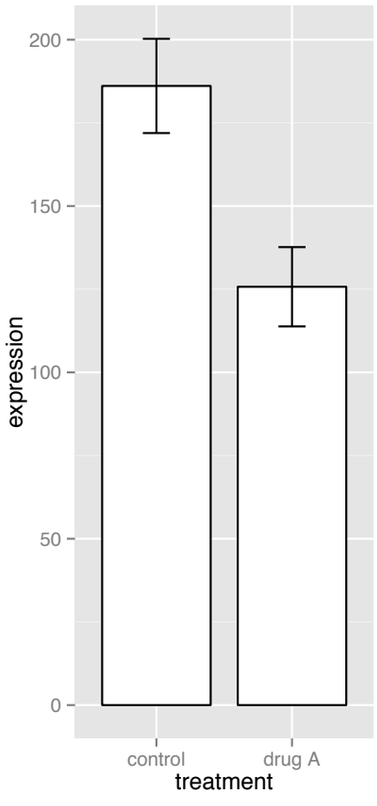
failure to repeat the result

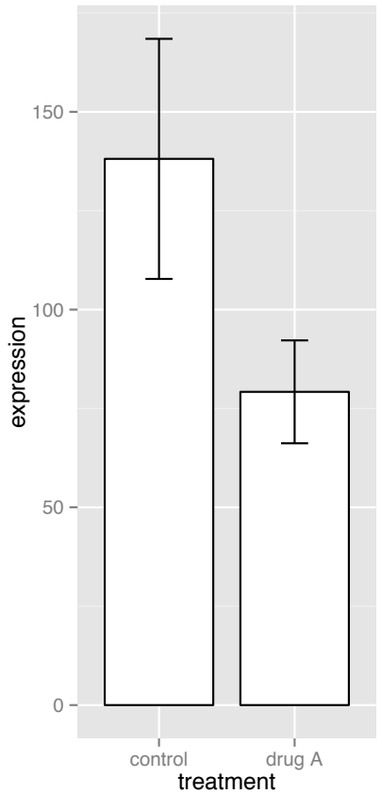


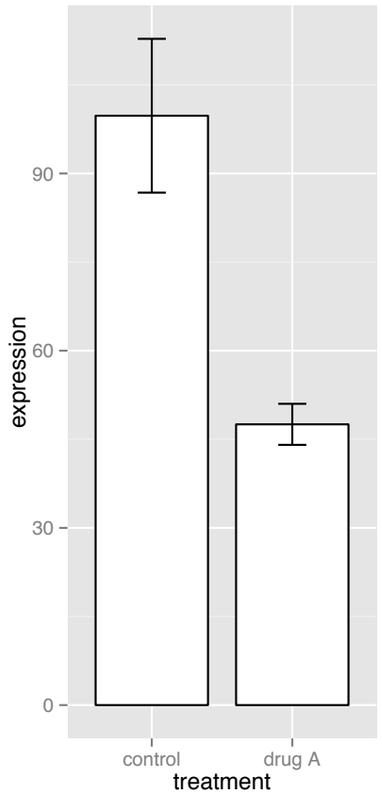
many repetitions

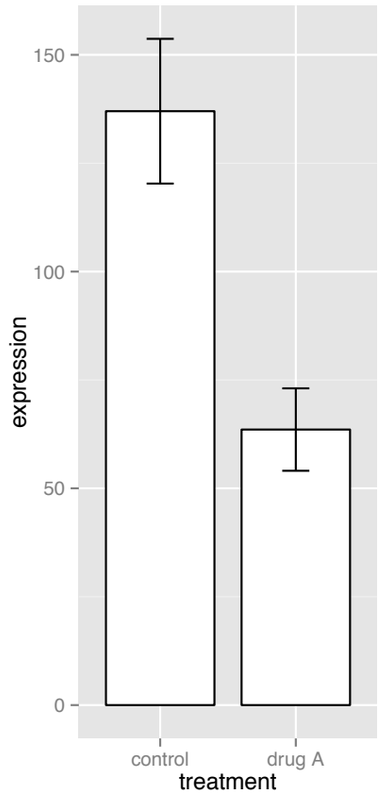


elimination of results





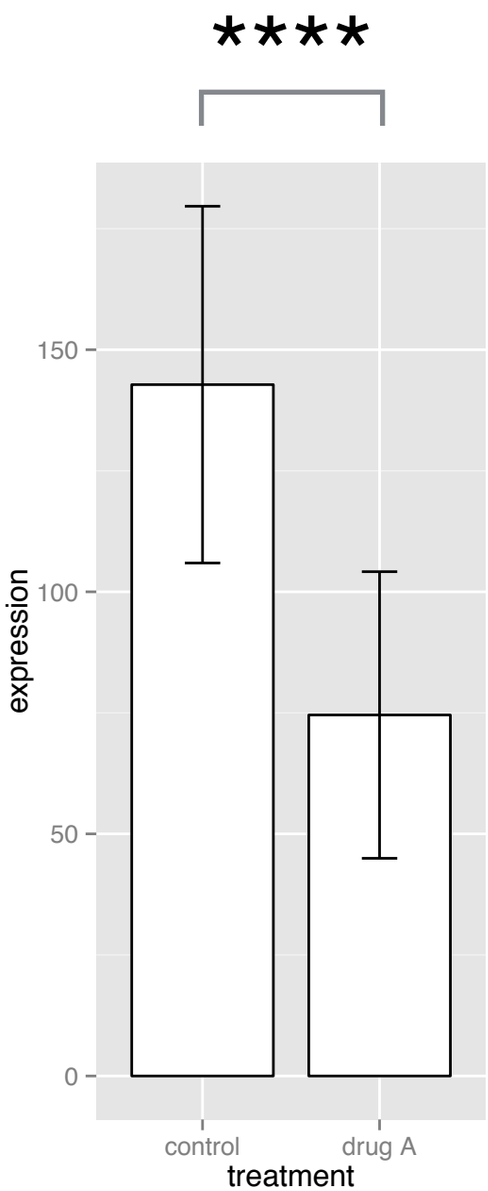




the statistical analysis

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

	Unpaired t test with equal SD	
1	Table Analyzed	Drug
2		
3	Column B	drug A
4	vs.	vs.
5	Column A	control
6		
7	Unpaired t test	
8	P value	< 0.0001
9	P value summary	****
10	Significantly different? (P < 0.05)	Yes
11	One- or two-tailed P value?	Two-tailed
12	t, df	t=5.592 df=28
13		
14	How big is the difference?	
15	Mean ± SEM of column A	142.8 ± 9.512, n=15
16	Mean ± SEM of column B	74.56 ± 7.642, n=15
17	Difference between means	-68.24 ± 12.20
18	95% confidence interval	-93.23 to -43.24
19	R squared	0.5276
20		
21	F test to compare variances	
22	F,DFn, Dfd	1.549, 14, 14
23	P value	0.4229
24	P value summary	ns
25	Significantly different? (P < 0.05)	No
26		



the manuscript

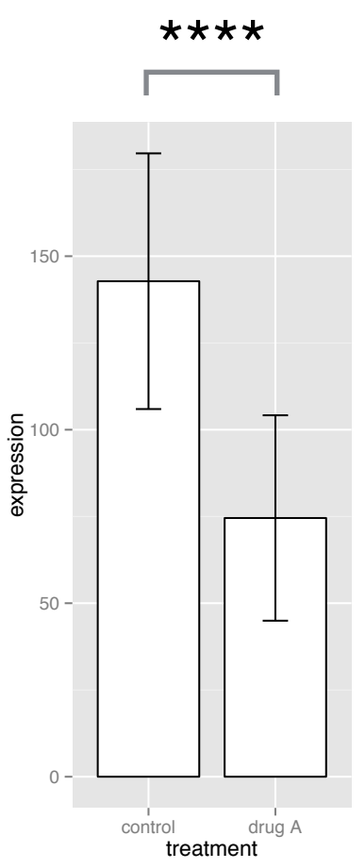


Fig.1: Drug A inhibits expression of gene x. RT-qPCR measurement of gene x transcript levels upon administration of a solvent control or 10 μM of drug A to proliferating XYZ cells for 24 hours.

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

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

.. gets published, original data deleted

the manuscript

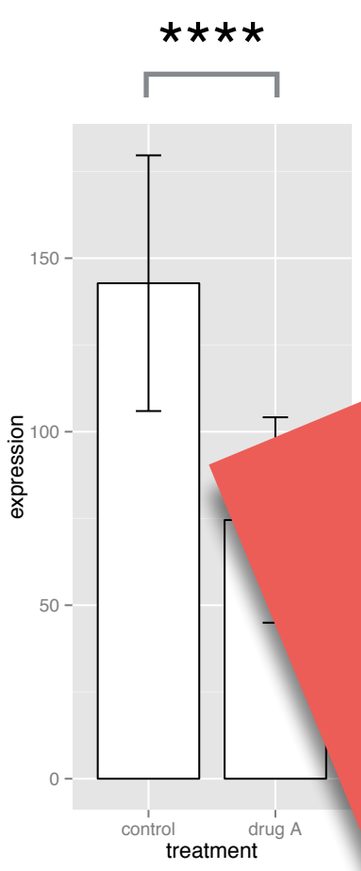


Fig.1: Drug A inhibits
expression of gene
qPCR measu
gene

Result

observed
nt effect
sion of
ht
reat

irreproducible

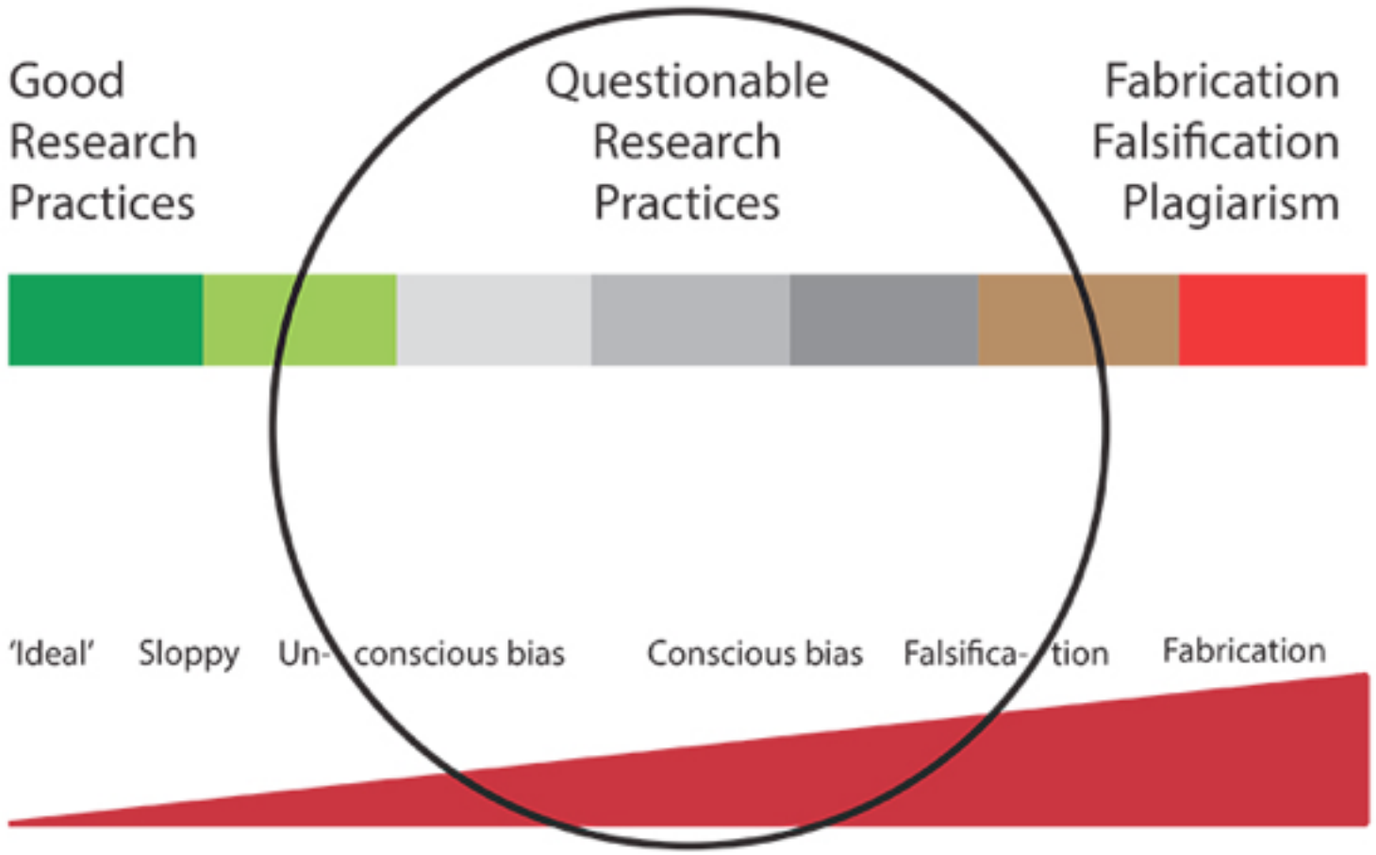
performed with Kit Q according
e[1]. Statistical analysis was done in
raphPad Prism.

.. gets published, original data deleted

Irreproducible because of..

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

QRP



http://www.vib.be/en/news/Pages/Research-misconduct---The-grey-area-of-Questionable-Research-Practices.aspx

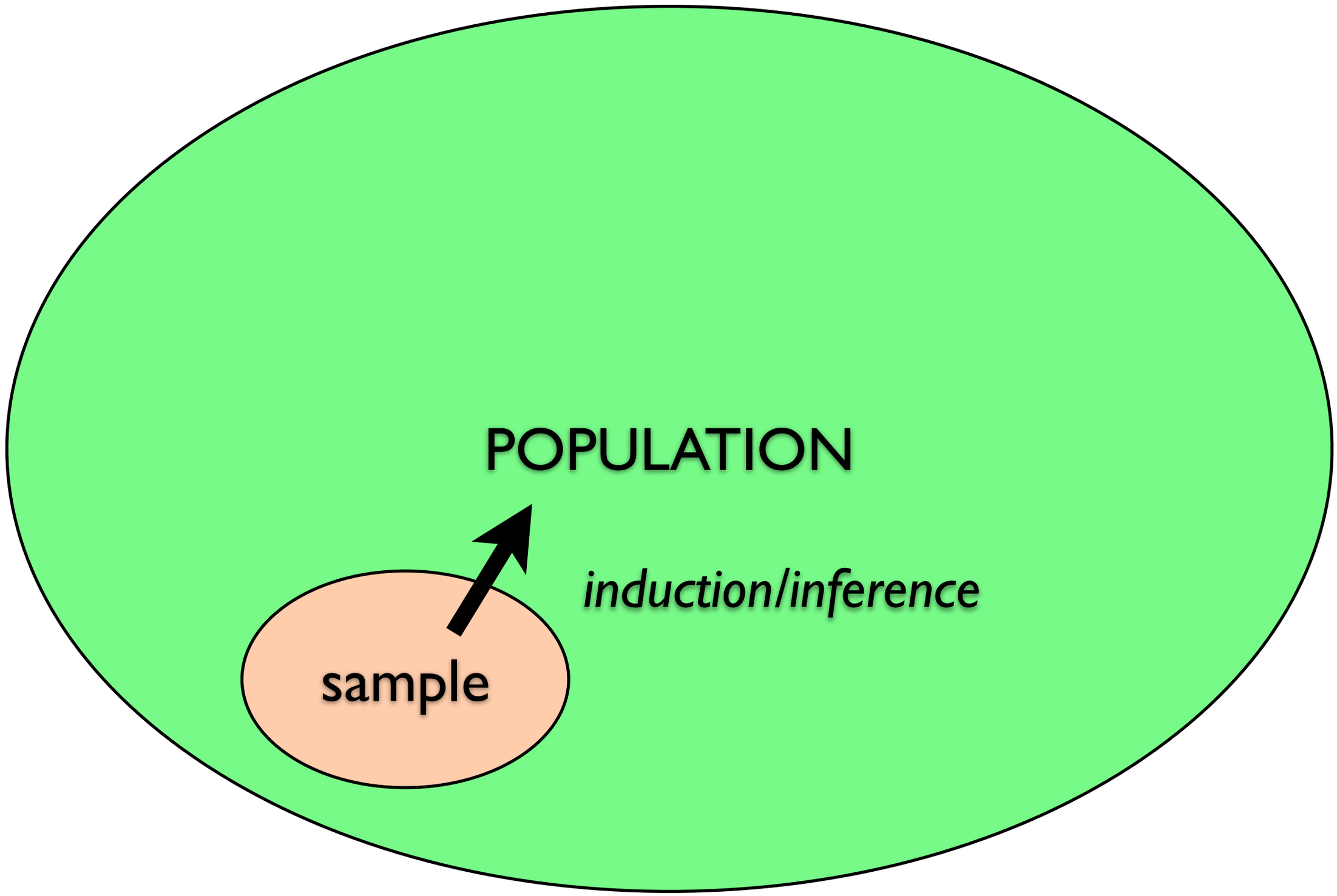
Examples of QRP:

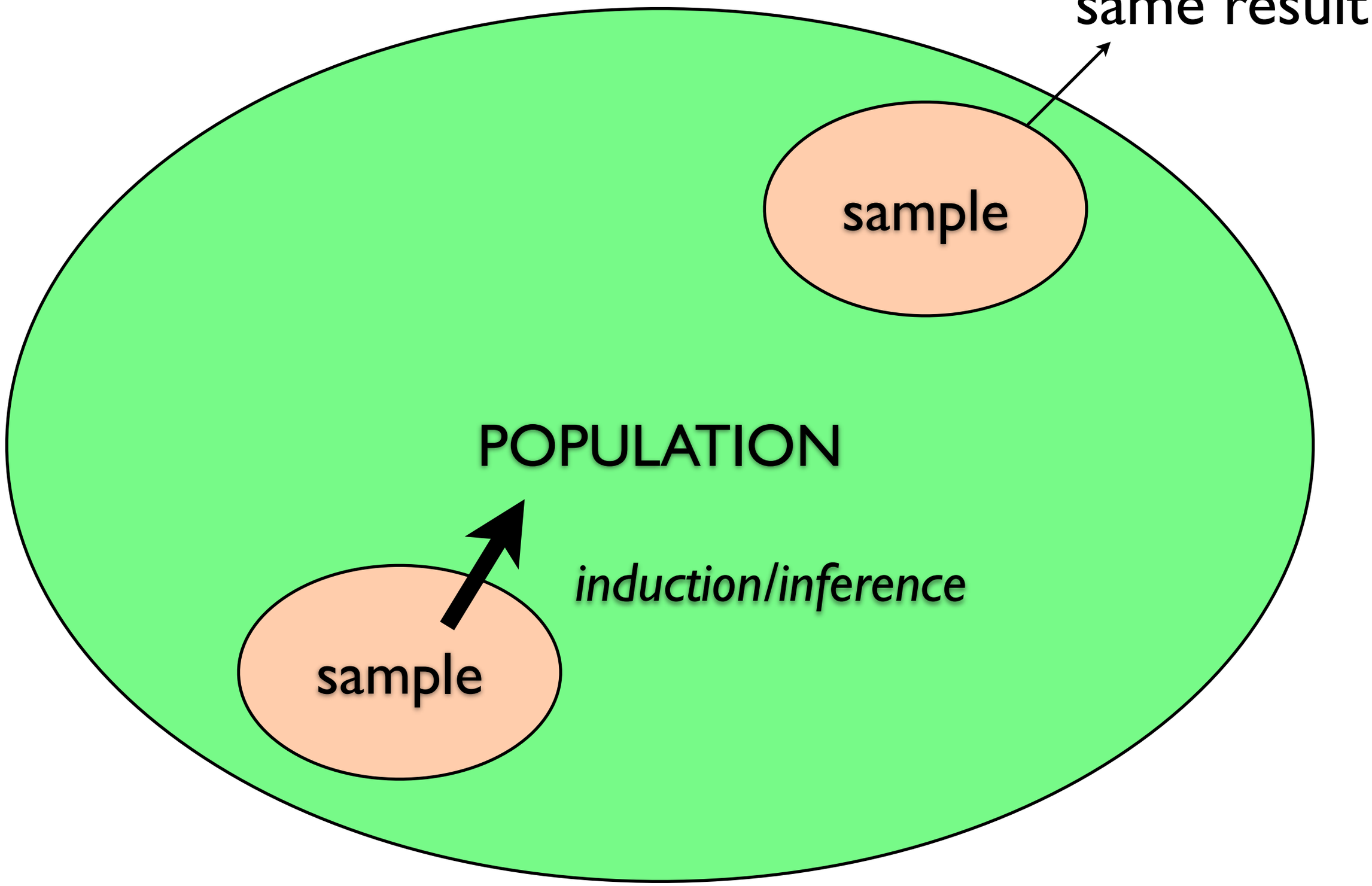
- Neglecting negative outcomes
- Using inappropriate statistics to support one's hypothesis
- Inappropriate research design
- Leaving out relevant controls
- Inappropriate re-use of controls
- Removal of 'outliers'
- Conscious bias
- Unethical experimentation
- Peer review abuse

key to reproducible research is the moment you start to think about reproducibility

How to generate
experimental results that are
valid in ***general***,
that will be ***reproducible***?

the sample - population
relationship


POPULATION
induction/inference
sample


same result
sample
POPULATION
induction/inference
sample

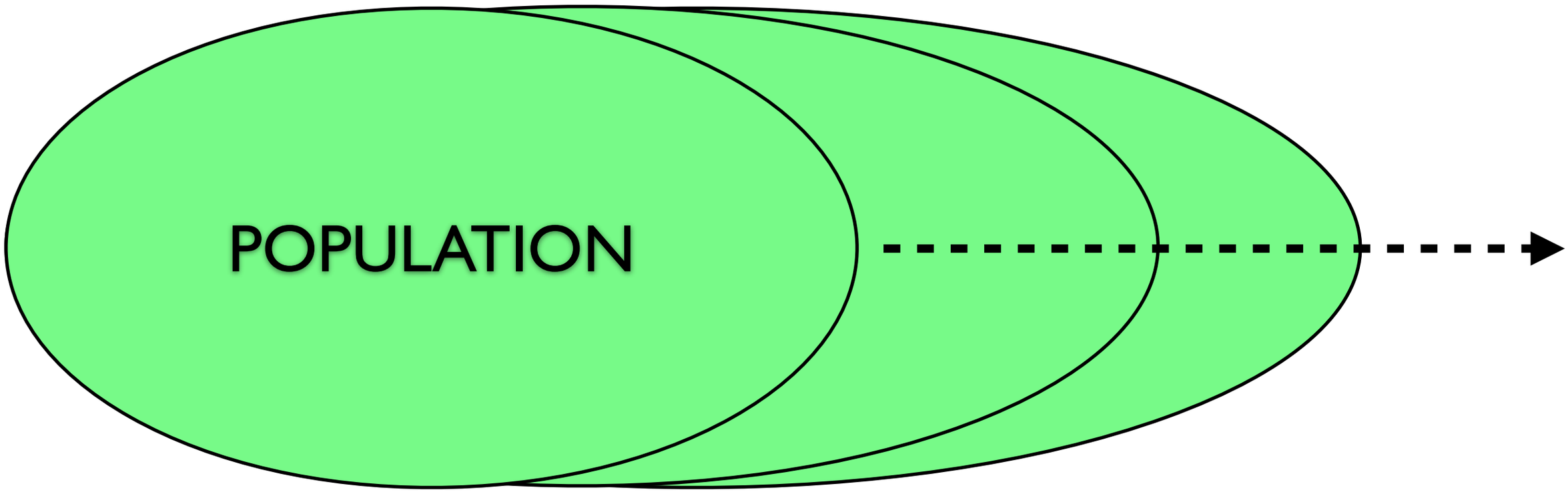
Replicability ⟷ Reproducibility

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



pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research

Replicability ← → Reproducibility

Reproduction of the origi… same protocol/reagents/tools

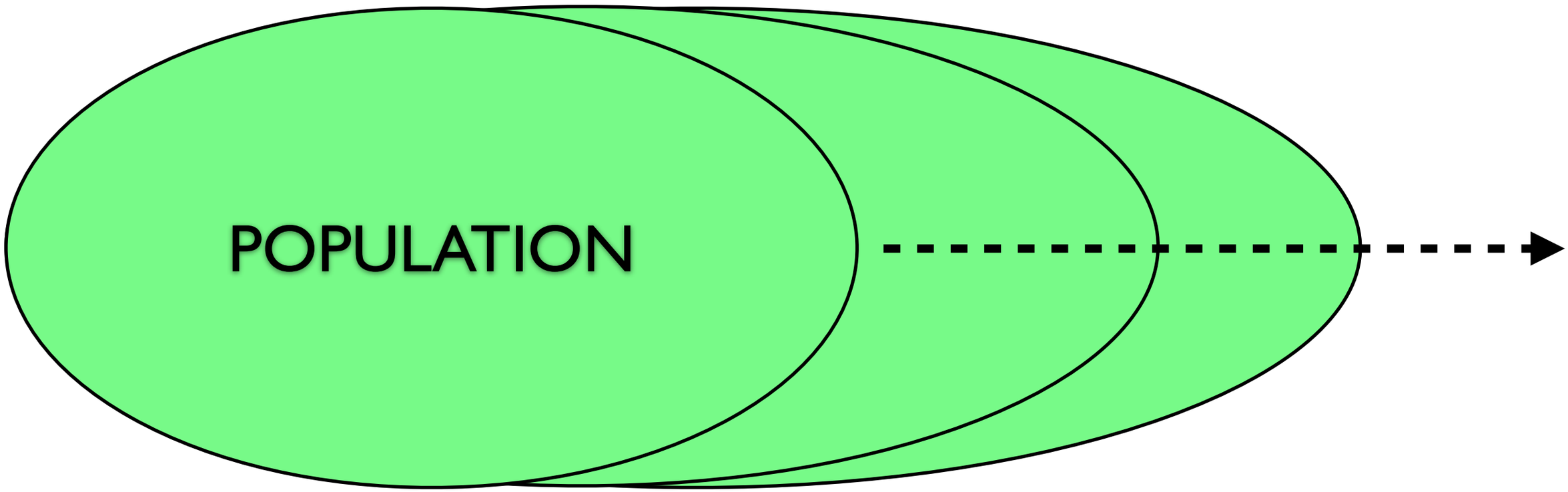
by the same person

by a different person in the lab

by a different person outside the lab

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

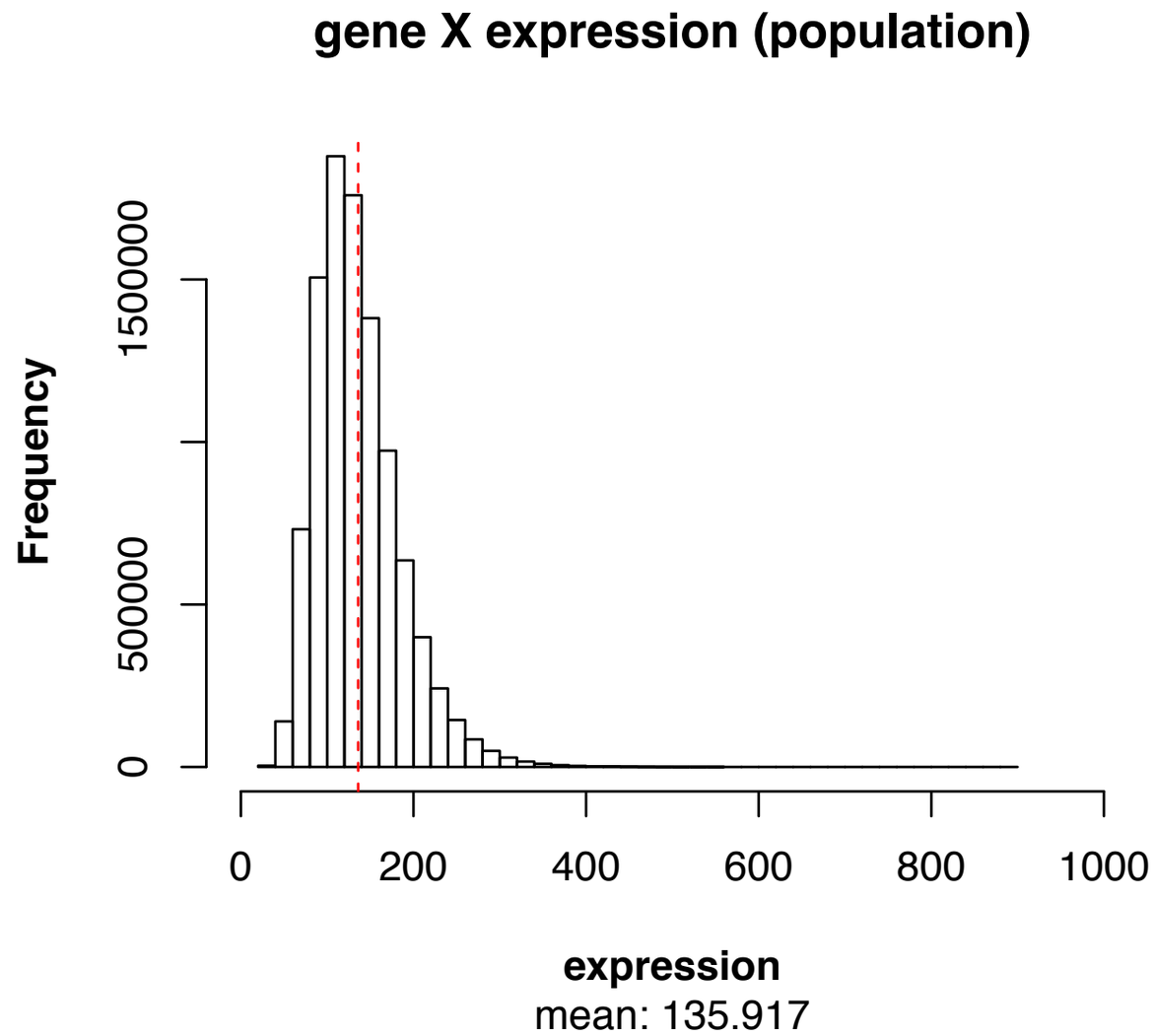
Reproduction just based on text description



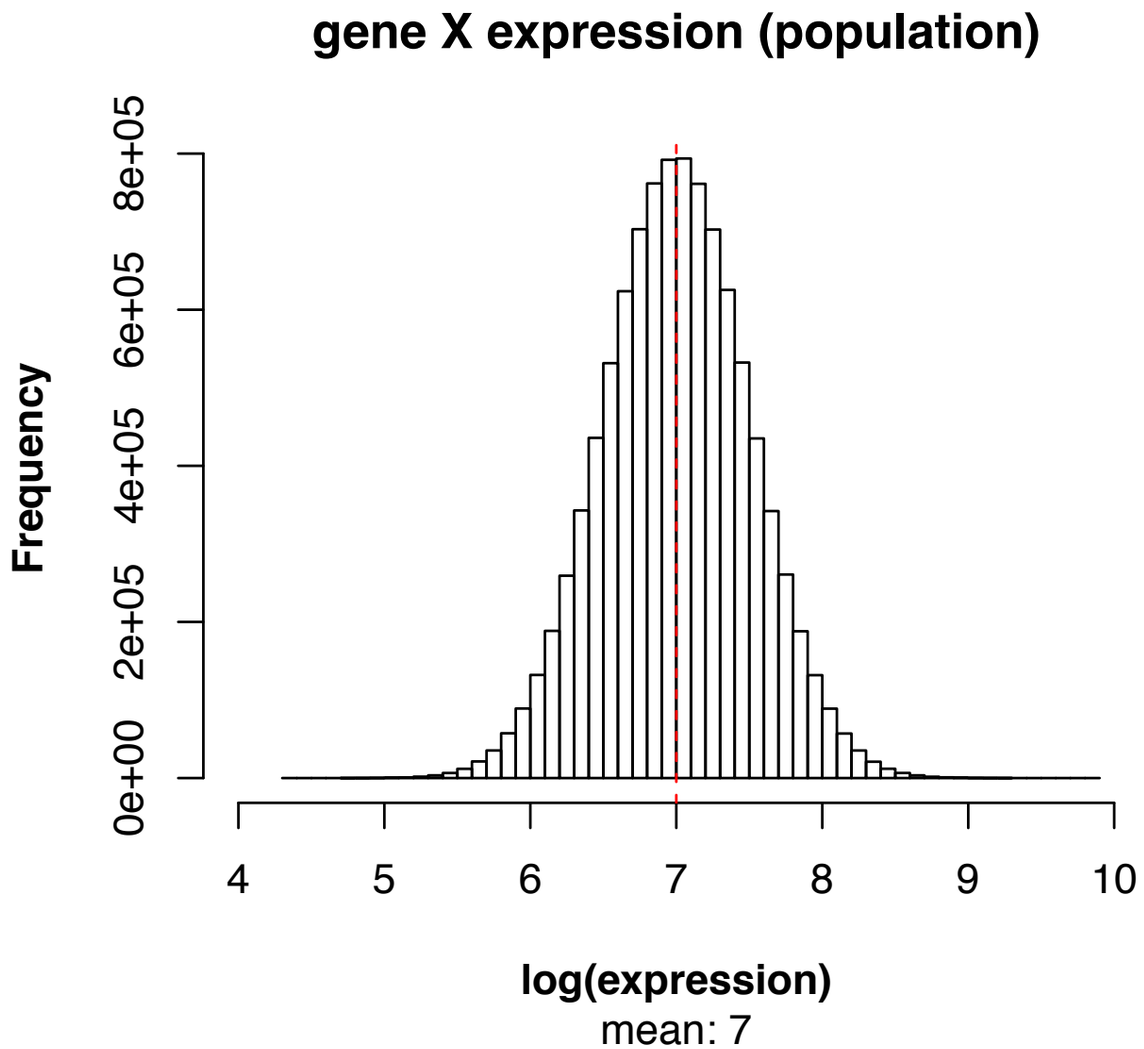
Inference

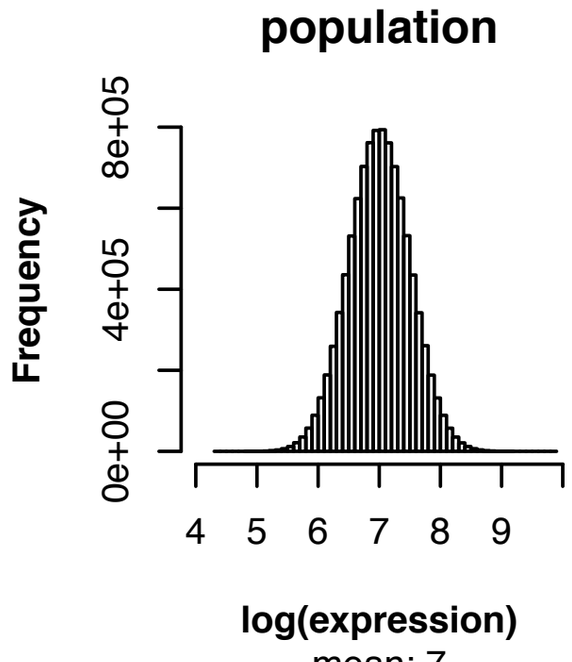
Howto

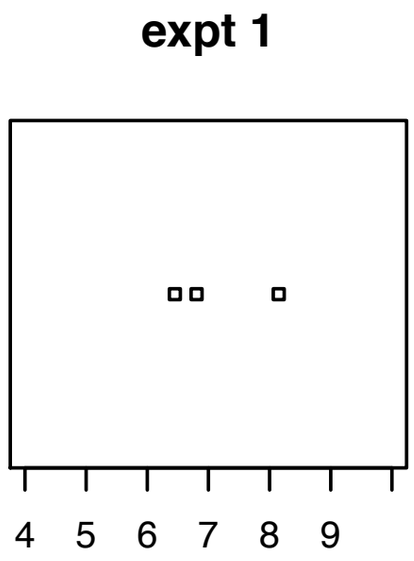
starting from the population

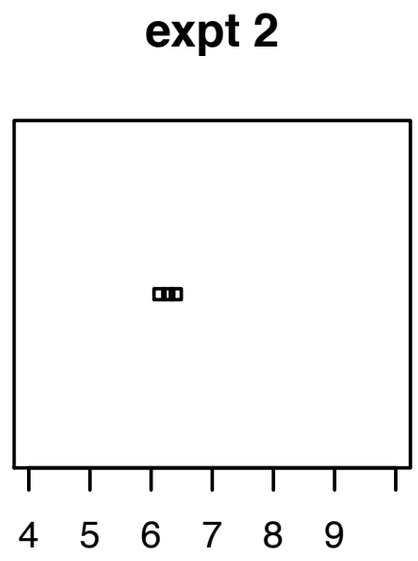


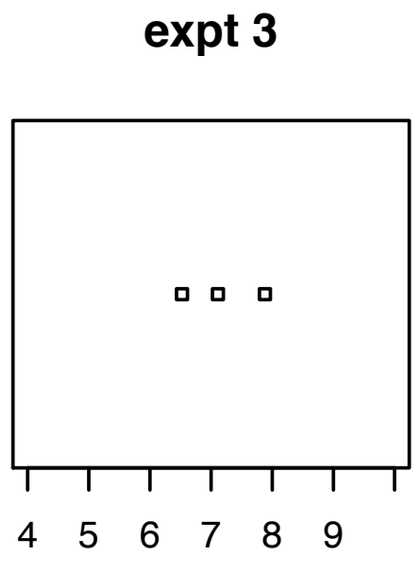
starting from the population

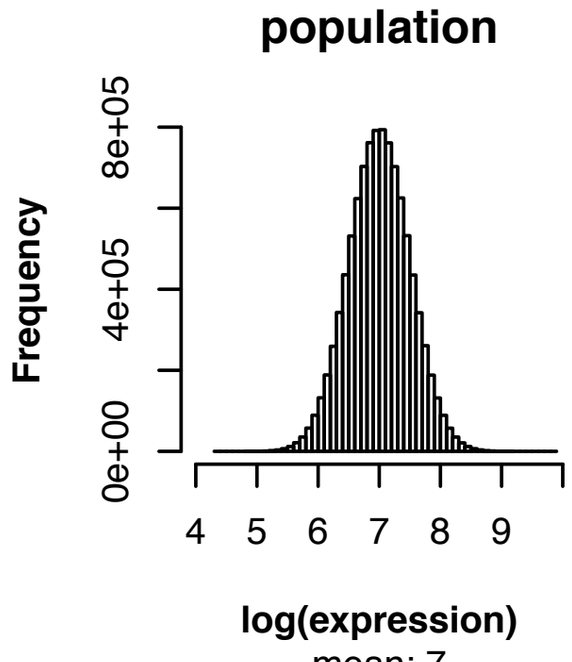


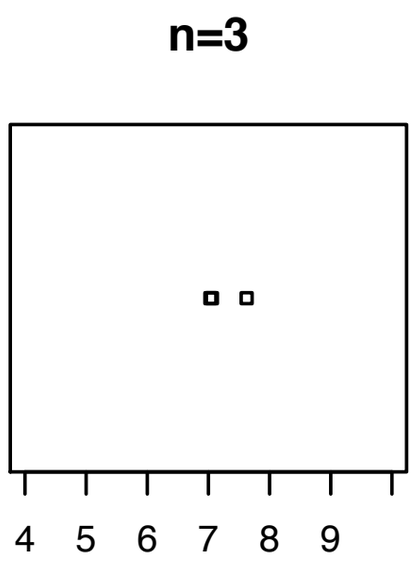

population
Frequency
0e+00
4e+05
8e+05
4
5
6
7
8
9
log(expression)
mean: 7

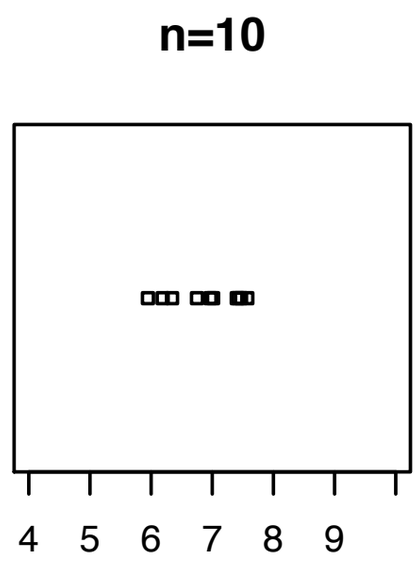

expt 1
4
5
6
7
8
9
log(expression)
mean: 7.135

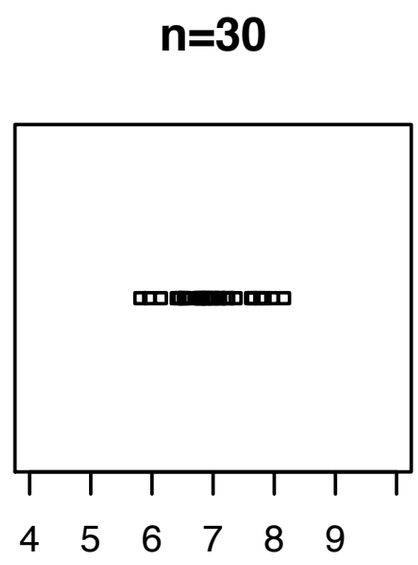

expt 2
4
5
6
7
8
9
log(expression)
mean: 6.274


expt 3
4
5
6
7
8
9
log(expression)
mean: 7.172


population
Frequency
0e+00
4e+05
8e+05
4 5 6 7 8 9
log(expression)
mean: 7


n=3
4 5 6 7 8 9
log(expression)
mean: 7.238


n=10
4 5 6 7 8 9
log(expression)
mean: 6.908

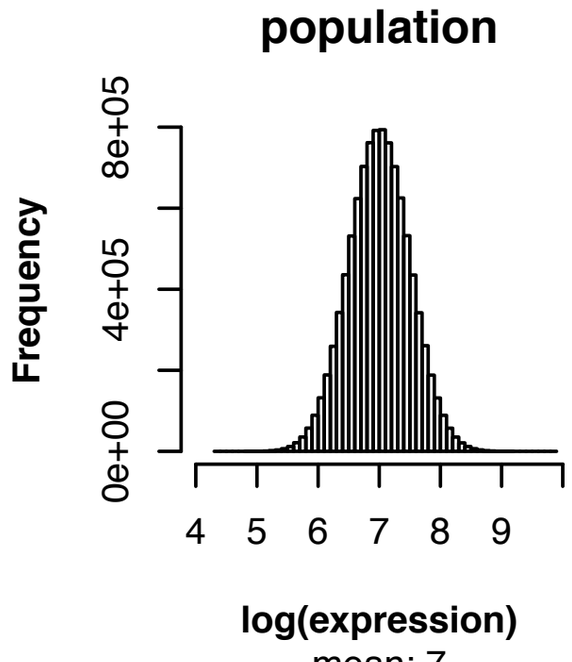

n=30
4 5 6 7 8 9
log(expression)
mean: 6.972

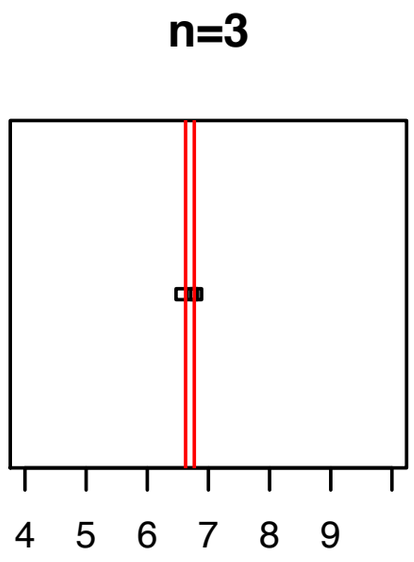
Standard Error (of the mean) - SEM

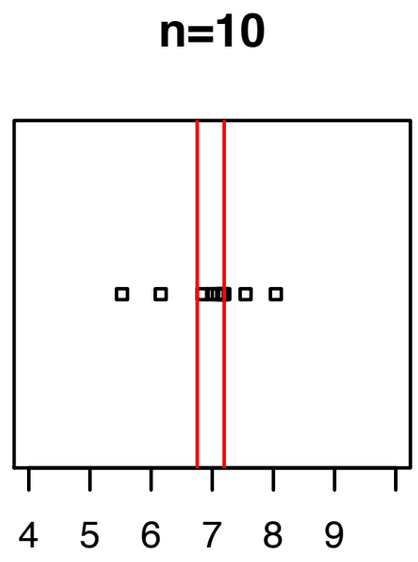
- The standard error of the mean (SEM) is the standard deviation of the sample mean estimate of a population mean.

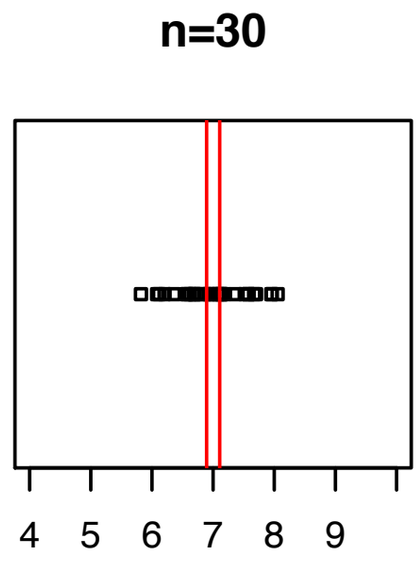
  SEM = standard deviation/square root(n)

- 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


population
Frequency
0e+00
4e+05
8e+05
4 5 6 7 8 9
log(expression)
mean: 7

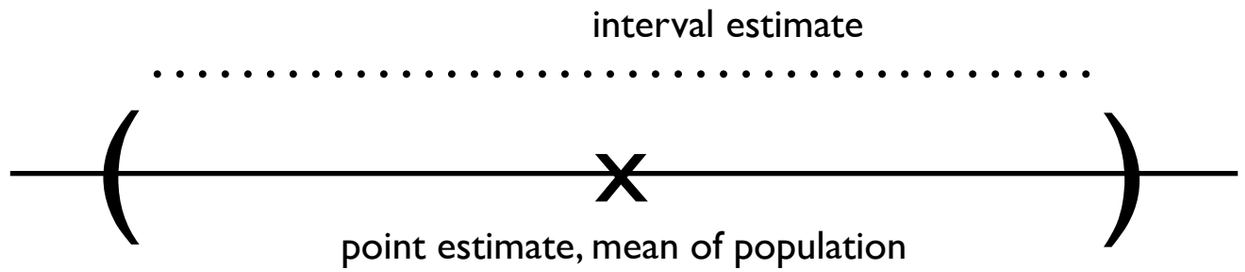

n=3
4 5 6 7 8 9
log(expression)
mean: 6.698


n=10
4 5 6 7 8 9
log(expression)
mean: 6.974

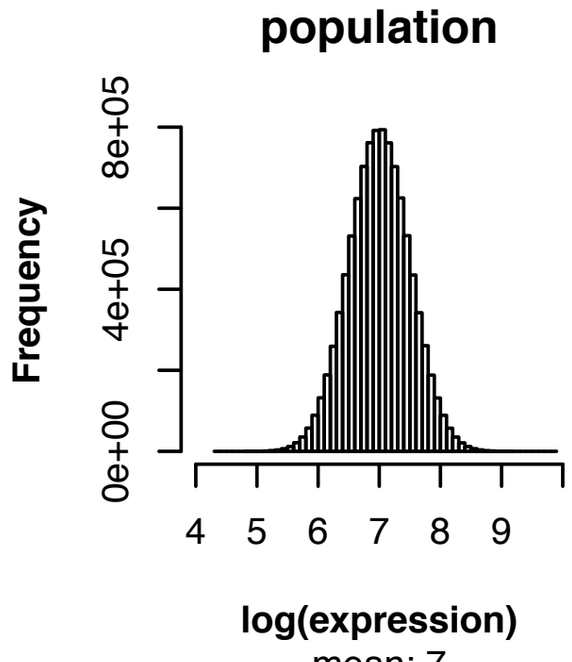

n=30
4 5 6 7 8 9
log(expression)
mean: 7.001

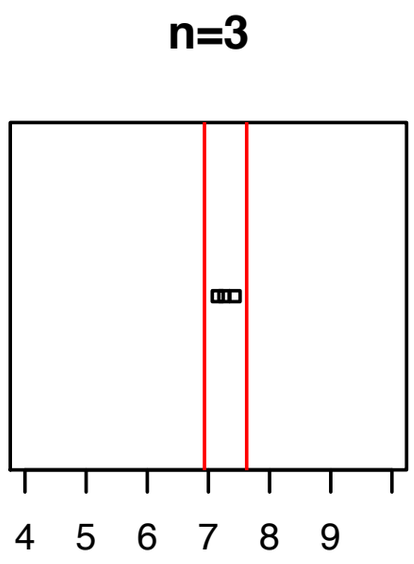
Confidence Intervals

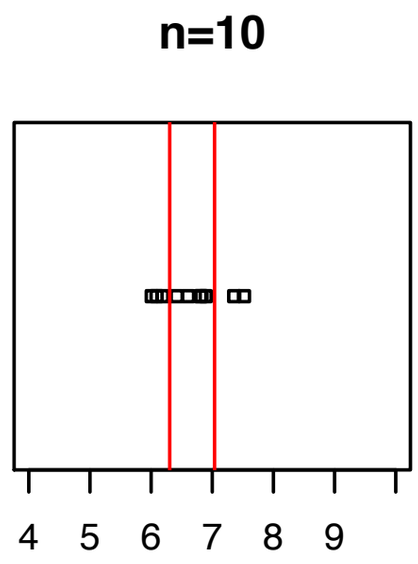
- 95%-confidence interval: An estimated interval which contains the „true value“ of a quantity with a probability of 95%.

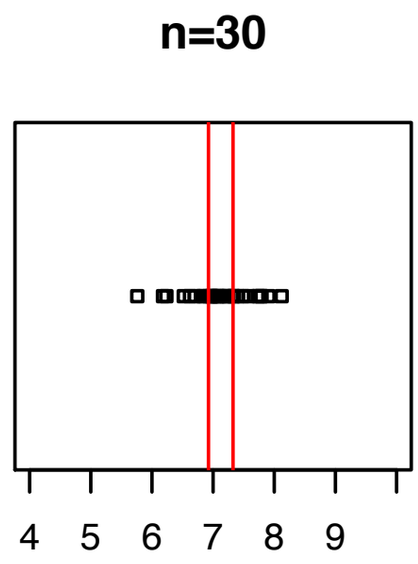


- (1– α)-confidence interval: An estimated interval which contains the „true value“ of a quantity with a probability of (1–α).
  1–α = confidence level, α = error probability


population
Frequency
0e+00
4e+05
8e+05
4 5 6 7 8 9
log(expression)
mean: 7


n=3
4 5 6 7 8 9
log(expression)
mean: 7.28


n=10
4 5 6 7 8 9
log(expression)
mean: 6.671


n=30
4 5 6 7 8 9
log(expression)
mean: 7.125

Different example

Someone asks: "how many dead cells are in your culture?"

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 reported?

Example taken from "Intuitive Biostatistics," H. Motulsky

Different example

Someone asks: "how many dead cells are in your culture?"

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 reported?

95% CI=0.02-0.13

Example taken from "Intuitive Biostatistics," H. Motulsky

Prerequisites for inference

- the sample has to be representative
- how is representativity achieved?
  - large sample number
  - independent sampling/random recruitment of samples

Technical replication

- The exact same sample is analysed multiple times.
- This addresses the variability of the analysis procedure (mass spectrometer, qPCR machine, pipetting errors etc.)
- Inference on the population aims however at the estimation of the biological variability. There is no interest convoluting this estimation with measurement errors.
- Technical variability should not be reported in the result of a biological experiment.
- Technically replicated measurements have to be averaged before inferential analysis.

Example

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

Summarise 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)

Example taken from “Intuitive Biostatistics,” H. Motulsky

	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

"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 (n=3). 95% CI (194.7;304.9)

Error Bars

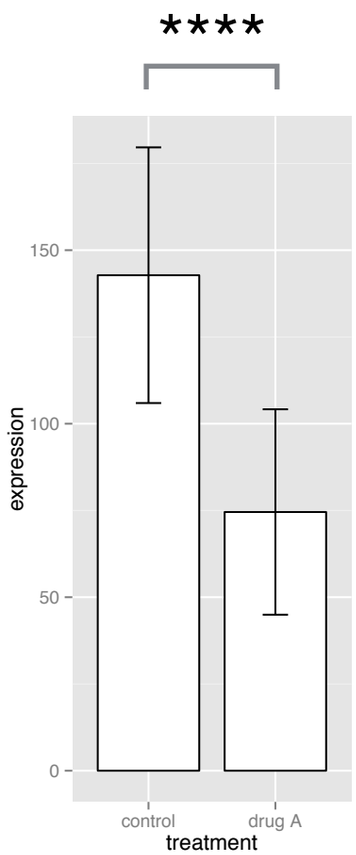


Fig.1: Drug A inhibits expression of gene x. RT-qPCR measurement of gene x transcript levels upon administration of a solvent control or 10 μM of drug A to proliferating XYZ cells for 24 hours.

- the type of error has to be reported (SD, SE…)
- n has to be reported
- errors (and statistics) should only be based on biological replication

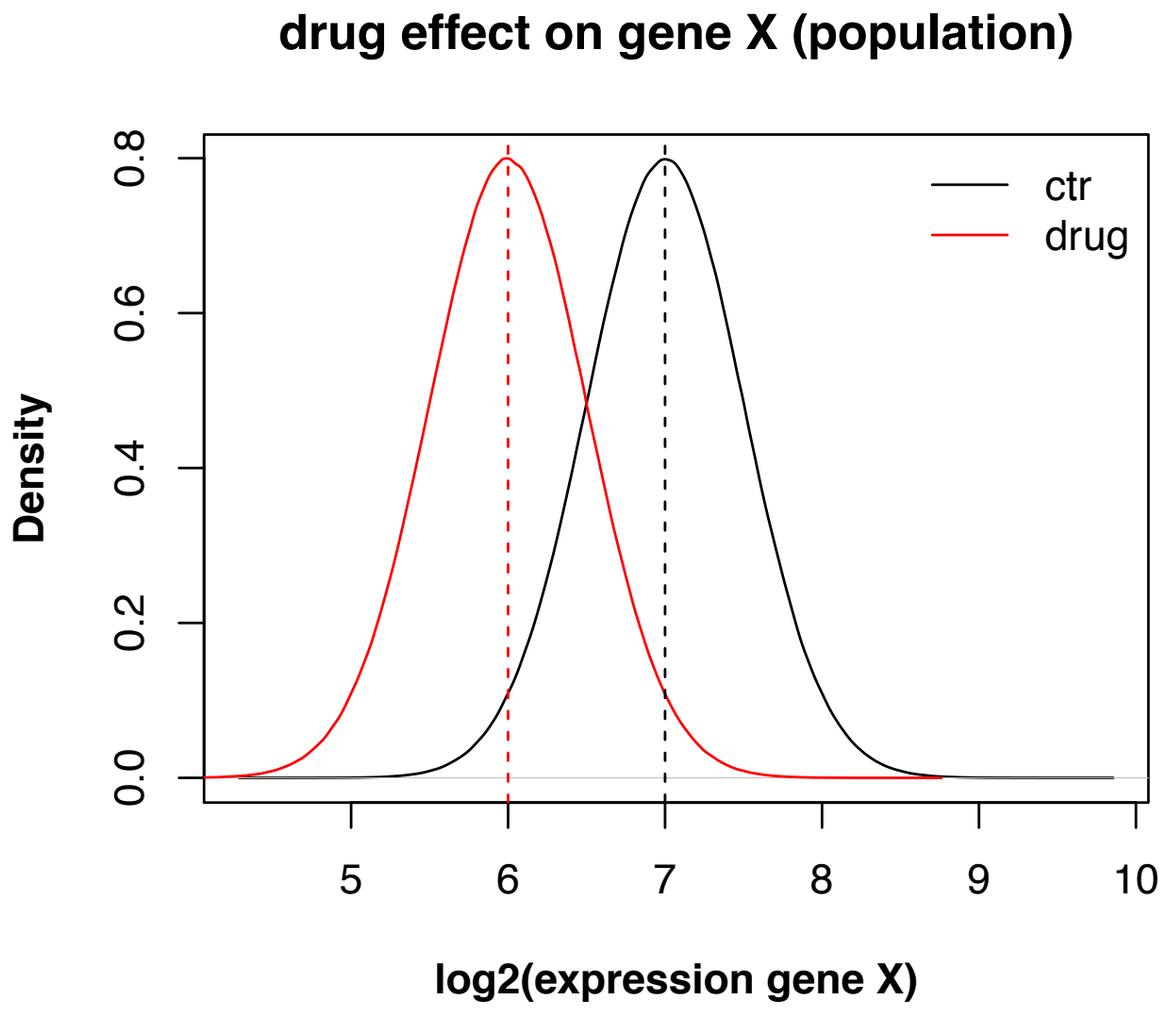
Error Bars

- 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.

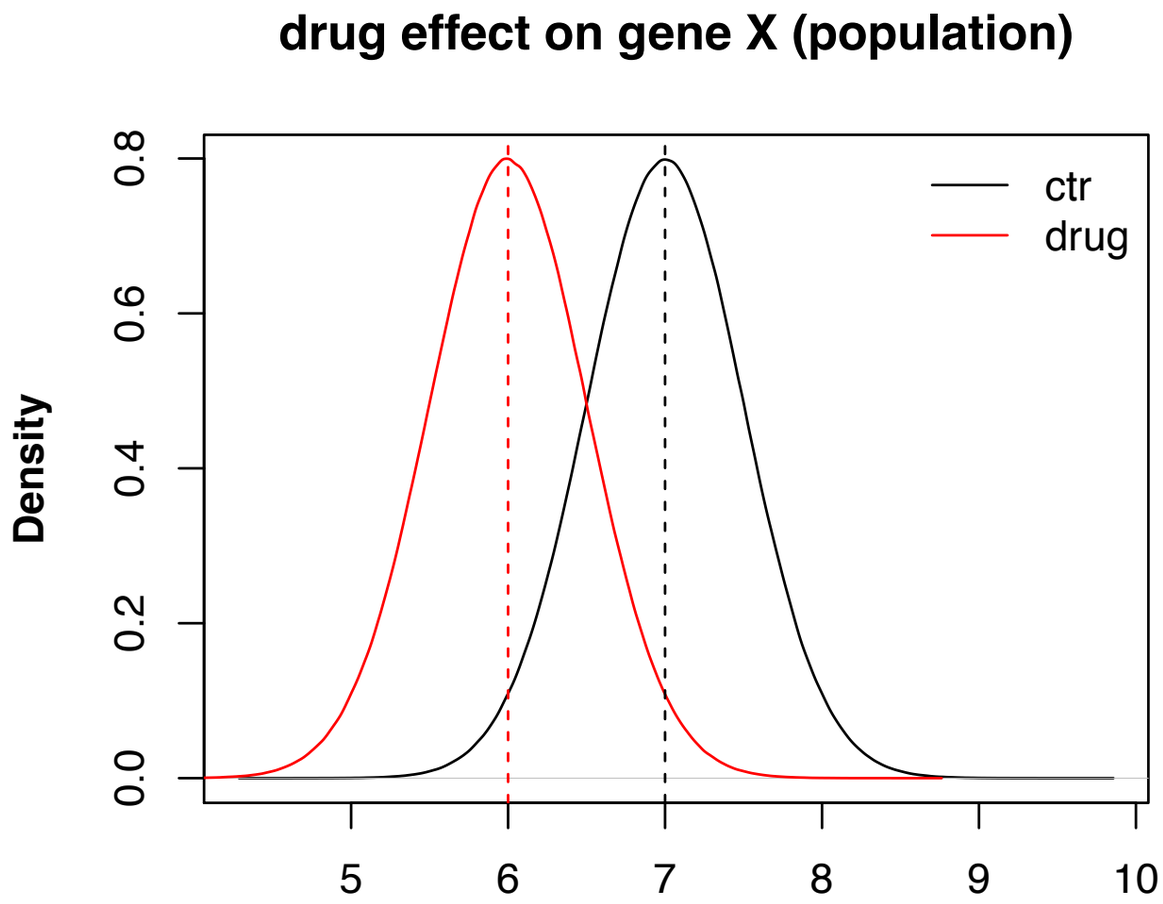
Error bars and significance

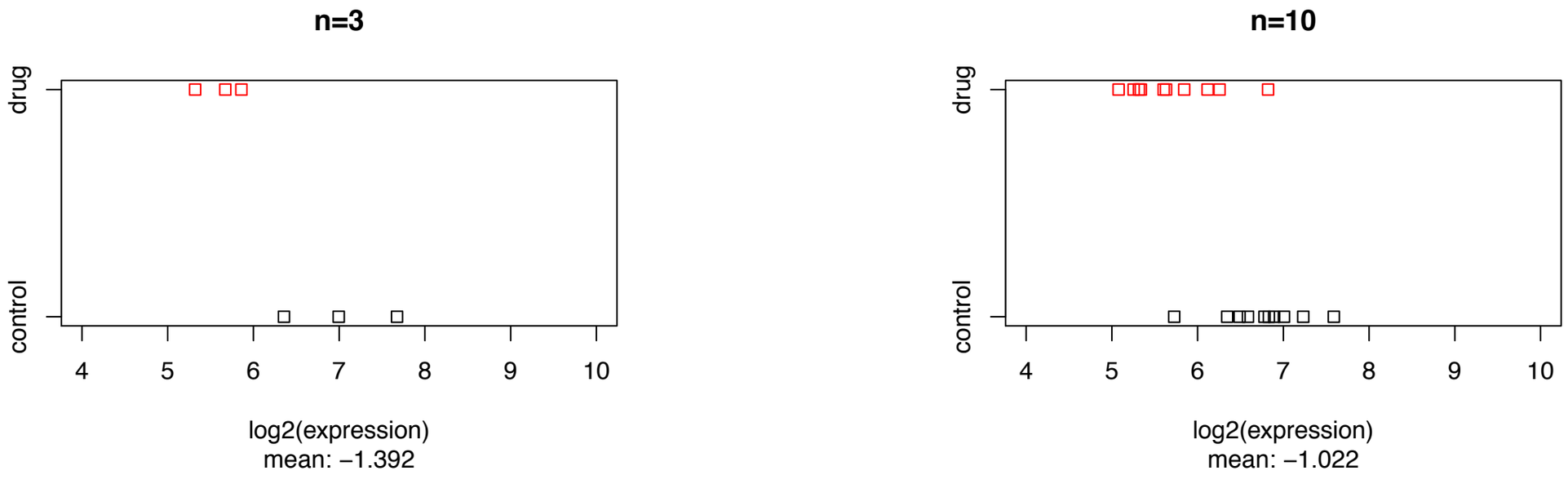
- The link between error bars and statistical significance is weaker than many wish to believe.
- But: if two SEM error bars overlap you can conclude that the difference is not statistically significant ($p>0.05$), but that the converse is not true.
- Some graphs and tables show the mean with the standard 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.

Looking at effects comparing population means

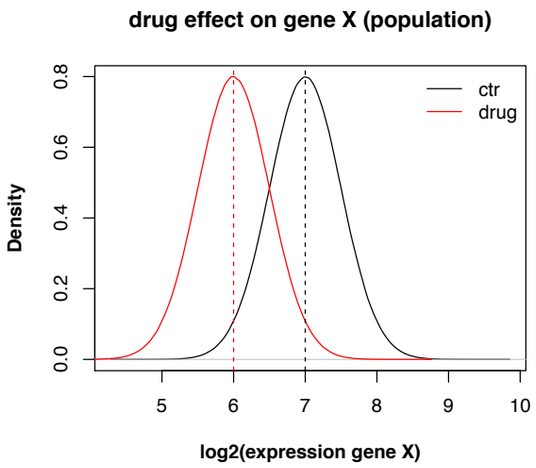


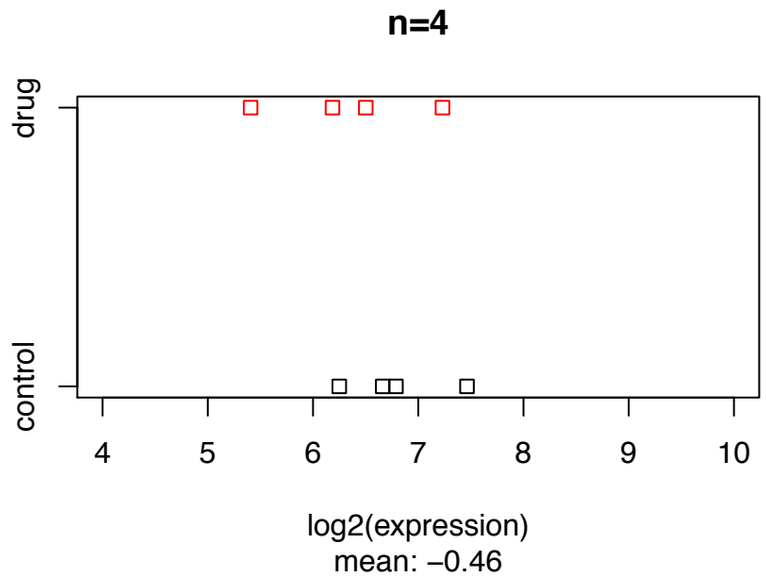
mean difference = mean(treatment)-mean(control)=-1 (0.5 in linear space)
= THE EFFECT

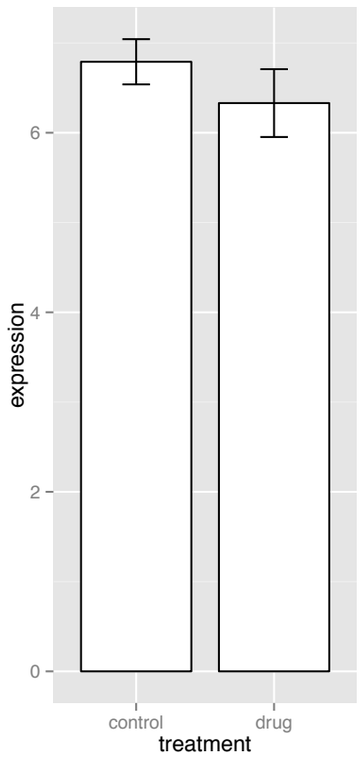

drug effect on gene X (population)
Density
0.0
0.2
0.4
0.6
0.8
ctr
drug
5
6
7
8
9
10


log2(expression gene X)
n=3
drug
control
4
5
6
7
8
9
10
log2(expression)
mean: −1.392
n=10
drug
control
4
5
6
7
8
9
10
log2(expression)
mean: −1.022

confidence interval of group means



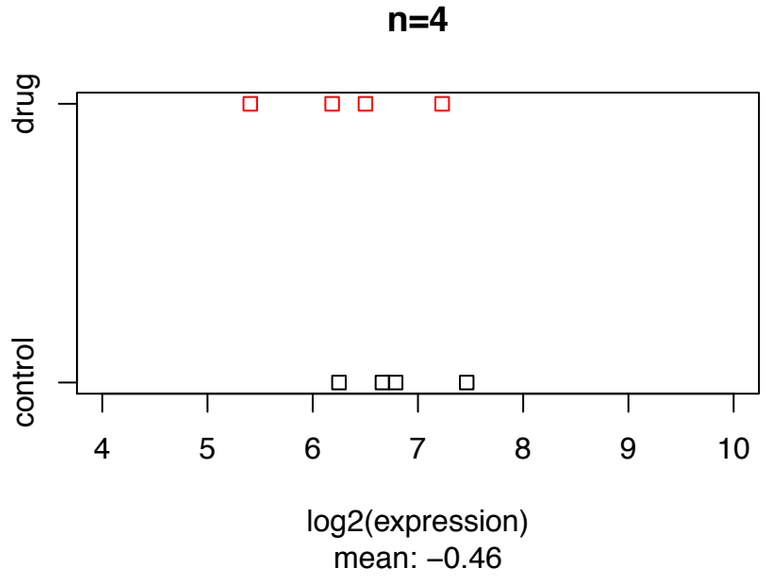


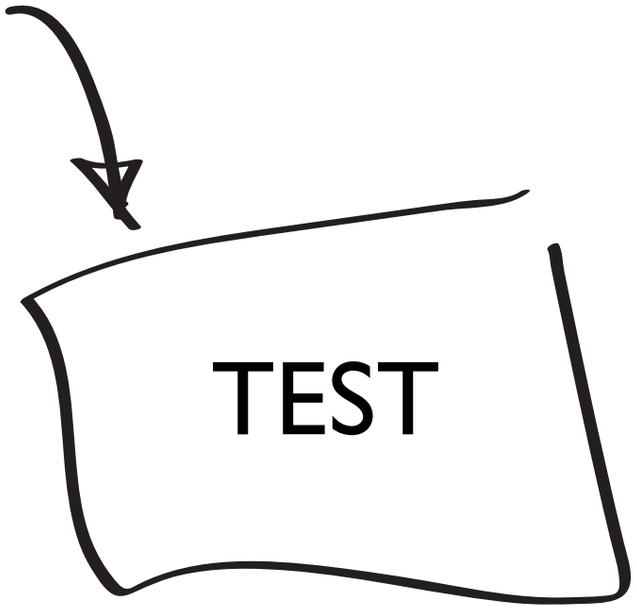


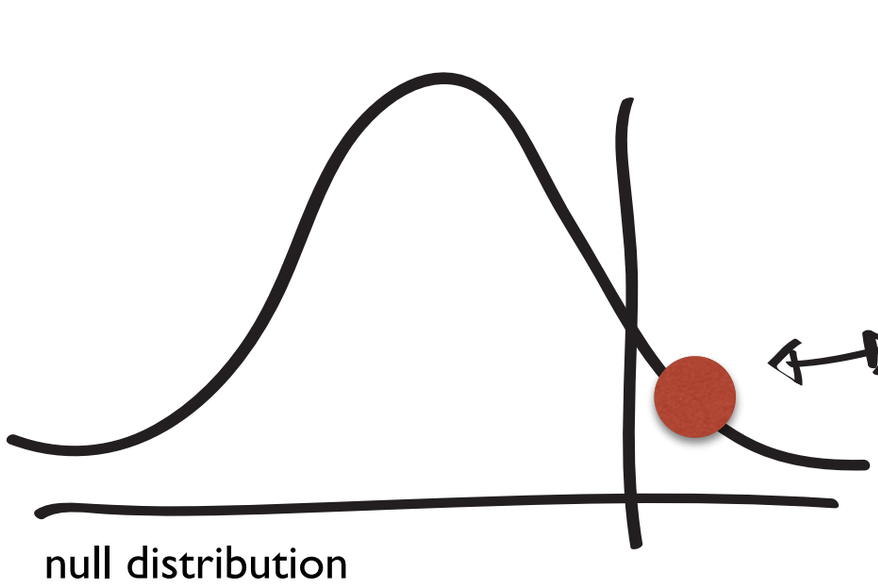
95 percent confidence interval:
(-1.61;0.69)

statistical (hypothesis) testing

Test for how likely an observed effect happened by chance (there was no effect)







value

p-value

significant!!??

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)

if $p < \alpha$

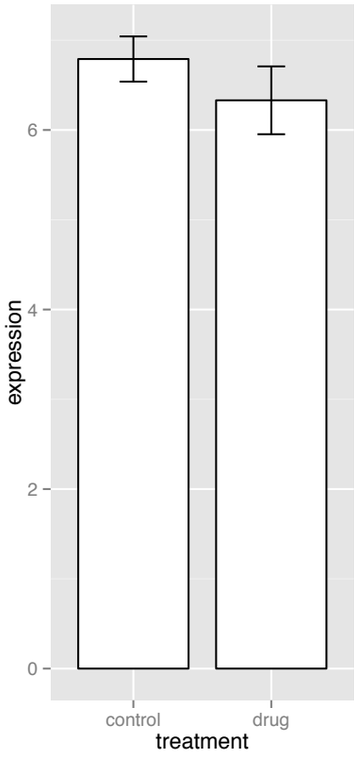
we reject the null hypothesis and call the result “significant”

prerequisite of statistical tests

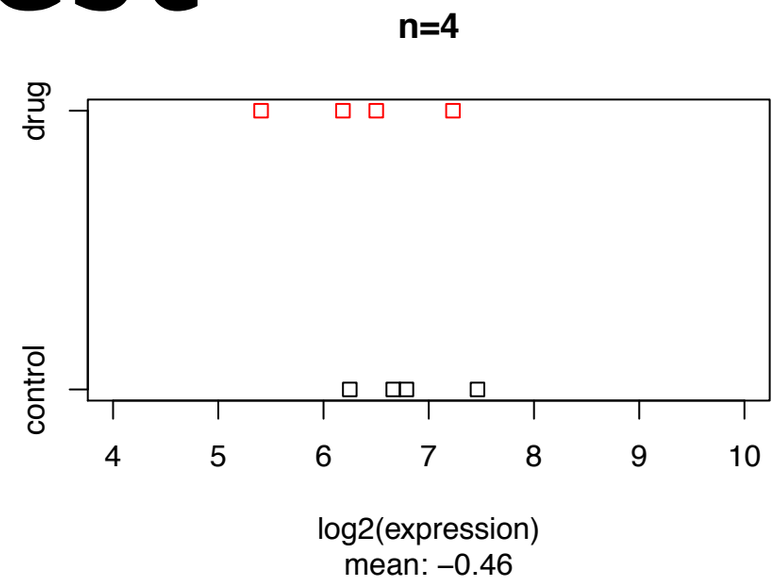
- formal requirements of the test procedures have to be met (distribution of the measurement values (normal, not-normal), equal variances across groups etc.)
- test parameters have to be set appropriately
- the decision for a test and its parameters has to be taken before data collection
- sampling has to be representative

two-sample t-test

- *null hypothesis*: there is no difference in the means of the measurements in the two groups (=drug has no effect)
- *alternative hypothesis*: there is a difference (=drug has an effect)
- *two-sided test*: the difference might be positive or negative
- *one-sided test*: the difference is either positive or negative
- t-test requires normal distribution of the measurements
- Student-t test requires equal variance



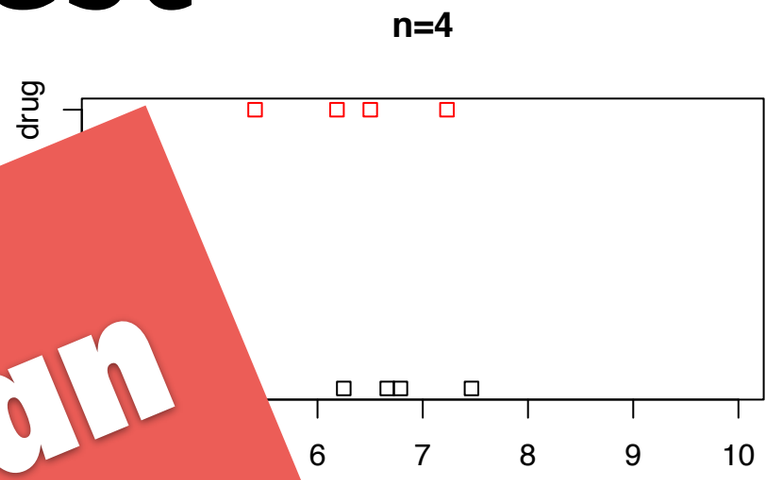
two-sample t-test



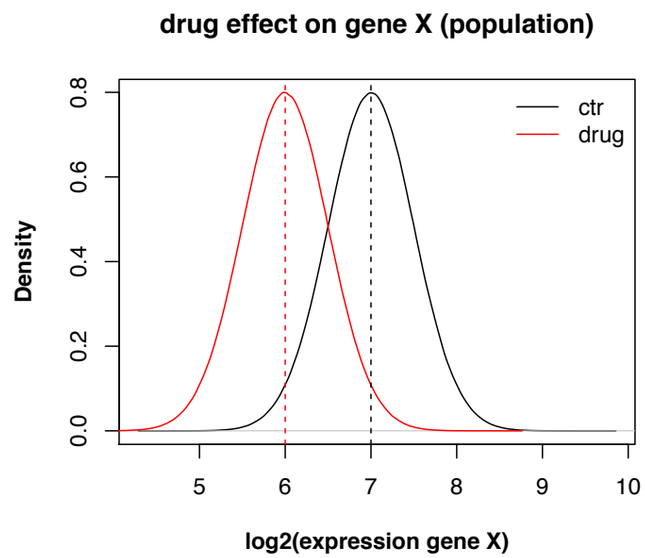

```
    Welch Two Sample t-test

data:  example.experiment
t = -1.0131, df = 5.228, p-value = 0.3556
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -1.6122931  0.6921616
sample estimates:
mean in group 1 mean in group 2
      6.330660        6.790726
```

two-sample t-test

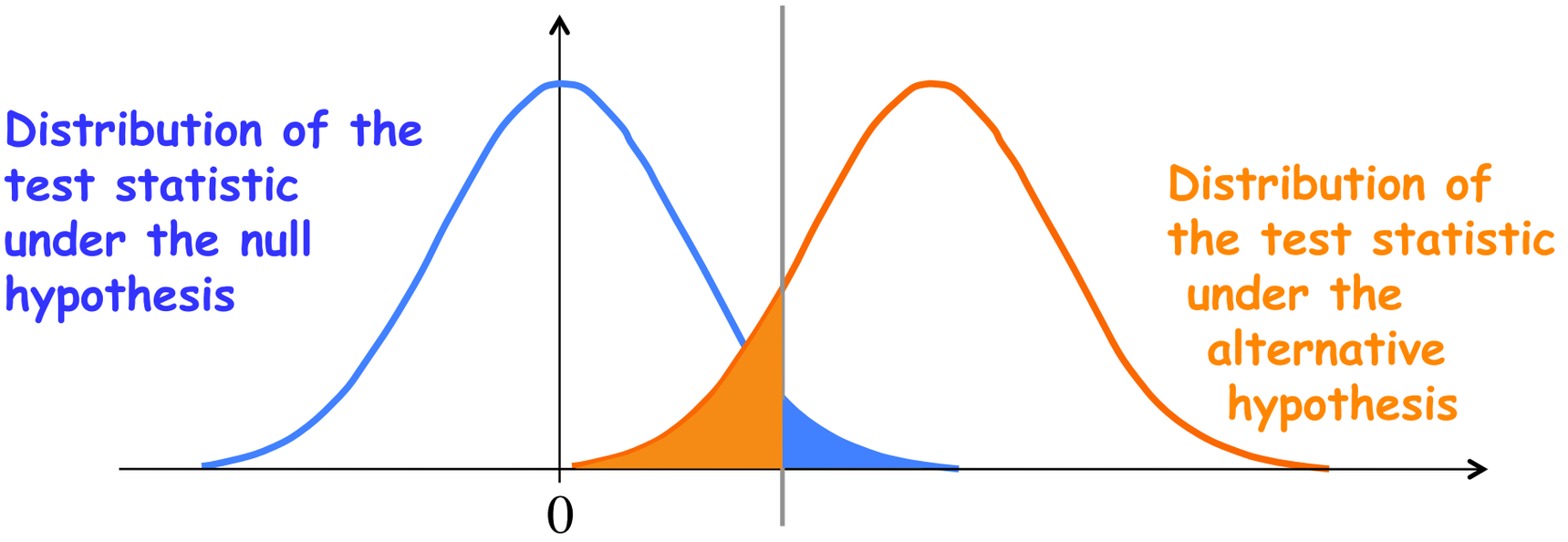
```
    Welch Two Sample t-test

data:  example.experiment
t = -1.0131, df =
alternative
equal
95 per
 -1.612
sample e
mean in gr
       6.33                 6.790726
```



but there was an effect!

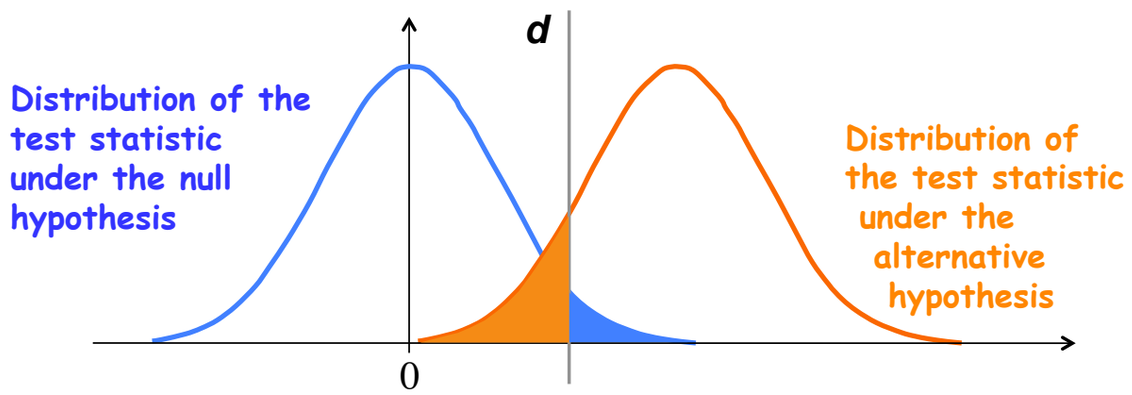


error of statistical tests



	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).
- 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.

power analysis

- sample size N
- effect size $d = \frac{\bar{x}_1 - \bar{x}_2}{s}.$
- α, significance level (0.05)
- power, 1-β (the probability of making a type II error) (typically set to 80% or 90%)
- specific for the test procedure
- can be performed before (interesting for the experimenter, search for N) and after (interesting for the interpreter, get the power) data collection

power analysis

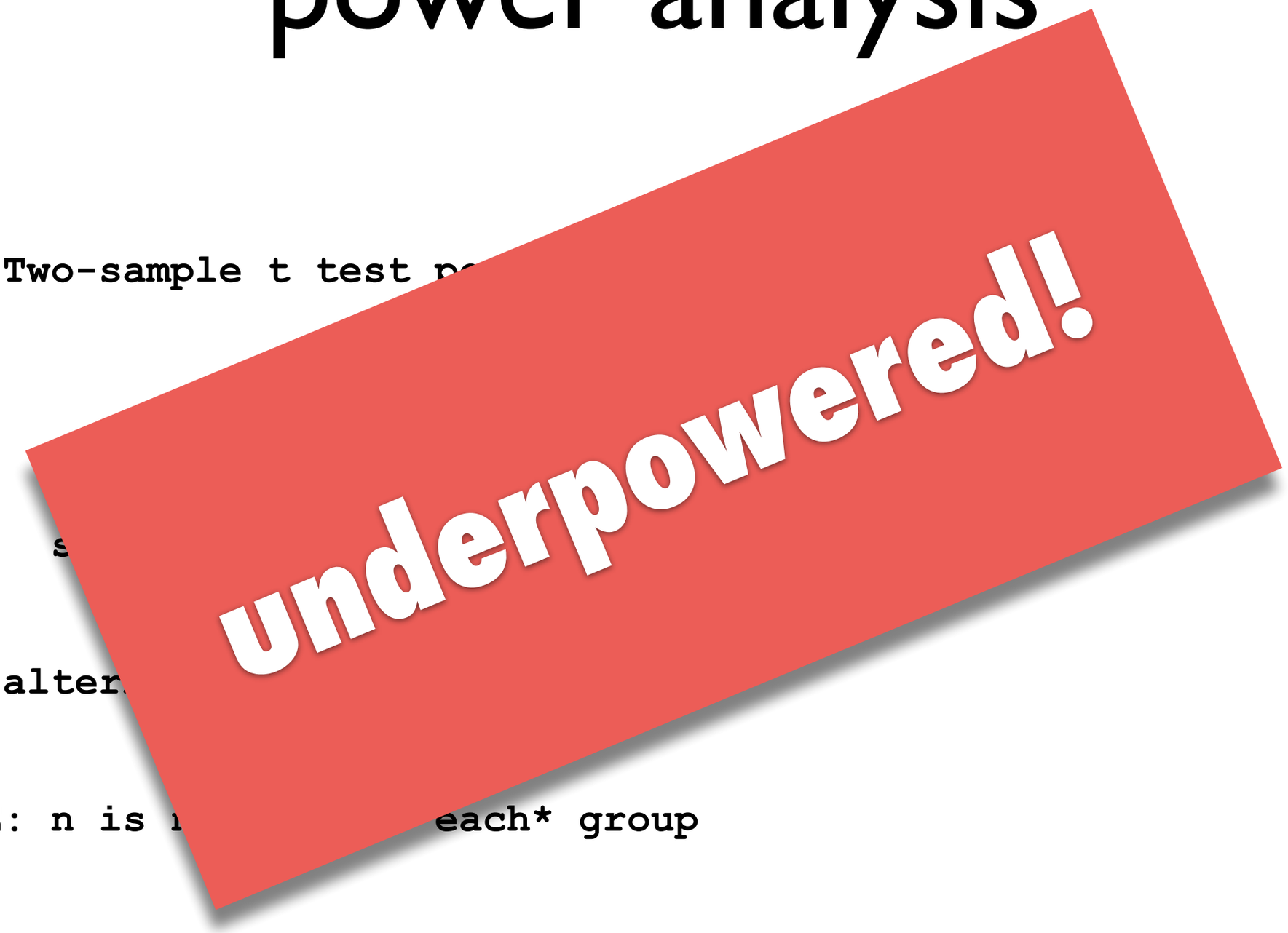
```
Two-sample t test power calculation

              n = 4
              d = 2
      sig.level = 0.05
          power =
    alternative = two.sided

NOTE: n is number in *each* group
```

power analysis

```
Two-sample t test power calculation

              n = 4
              d = 2
      sig.level = 0.05
          power = 0.6568759
    alternative = two.sided

NOTE: n is number in *each* group
```

power analysis

```
Two-sample t test po

          s

     alter

NOTE: n is                each* group
```



underpowered studies

- have a low sensitivity
- correlate with irreproducibility

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2

power analysis - obtaining N

```
Two-sample t test power calculation

              n =
              d = 2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

power analysis - obtaining N

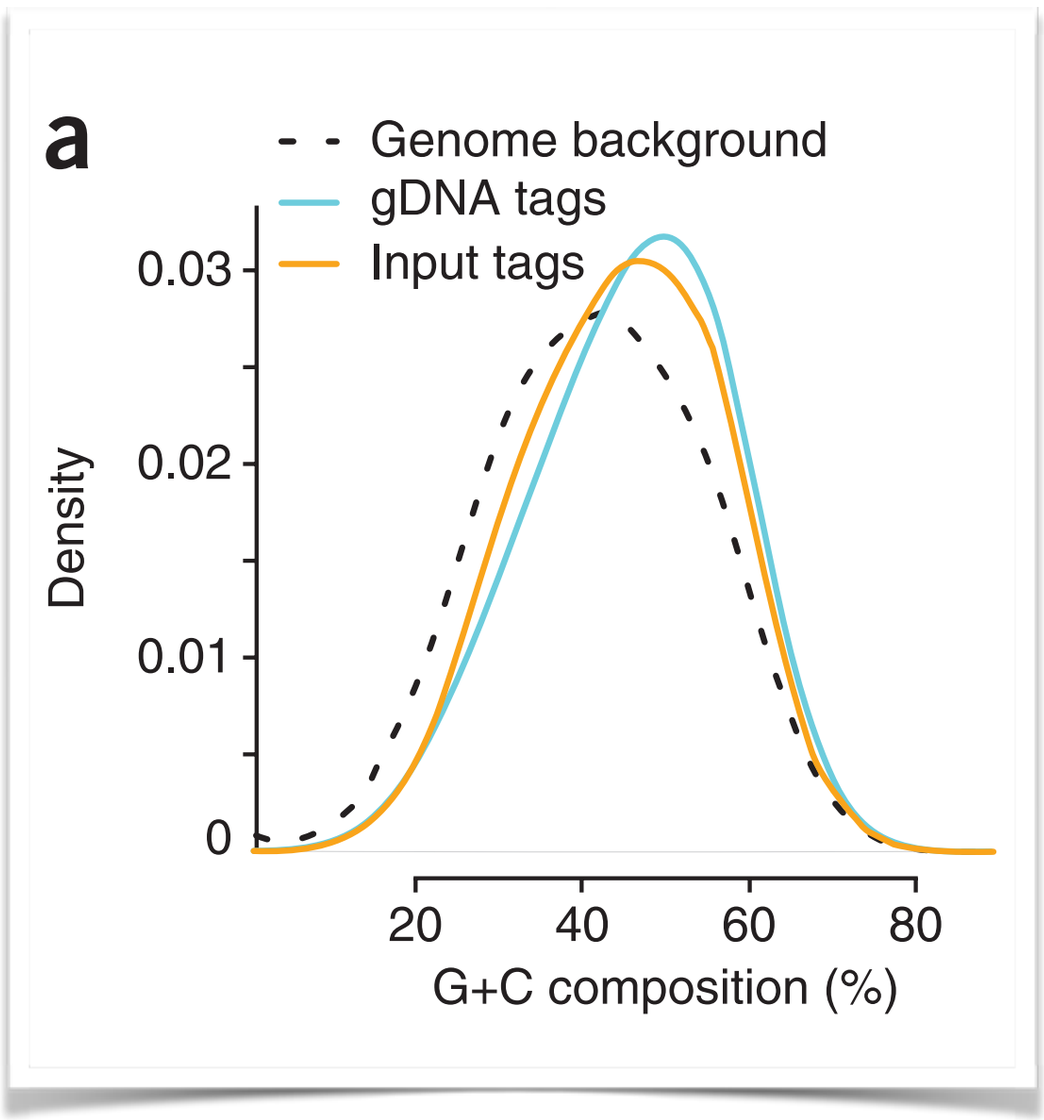
```
     Two-sample t test power calculation

              n = 5.090002
              d = 2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

overpower!

```
     Two-sample t test power calculation

              n = 10000
              d =
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

overpower!

```
     Two-sample t test power calculation

              n = 10000
              d = 0.03962599
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```



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

p-value > 0.05 does not prove equality!

summary

statistical inferences require

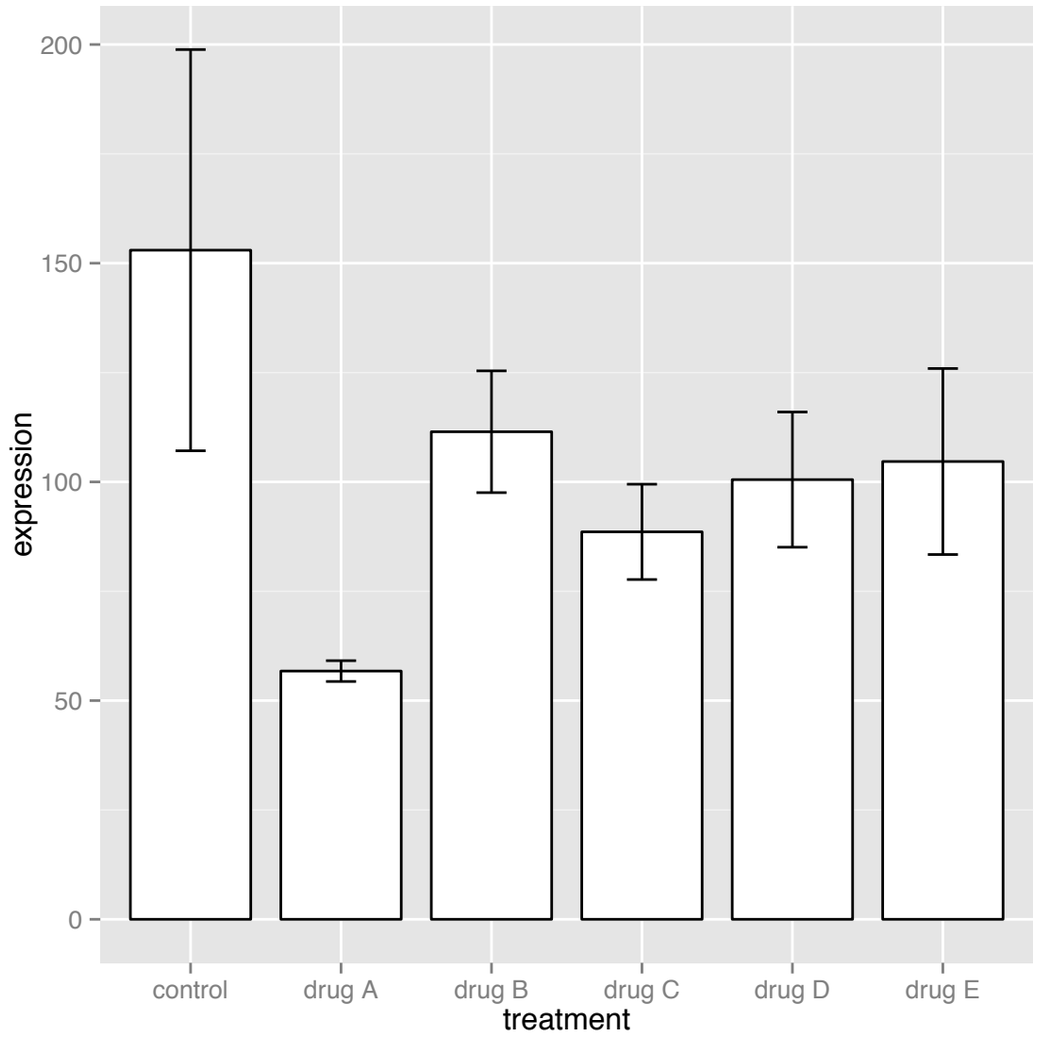
- fulfilment of prerequisites for statistical testing
- the test to be adequately powered

Back to the beginning

some thoughts before start pipetting

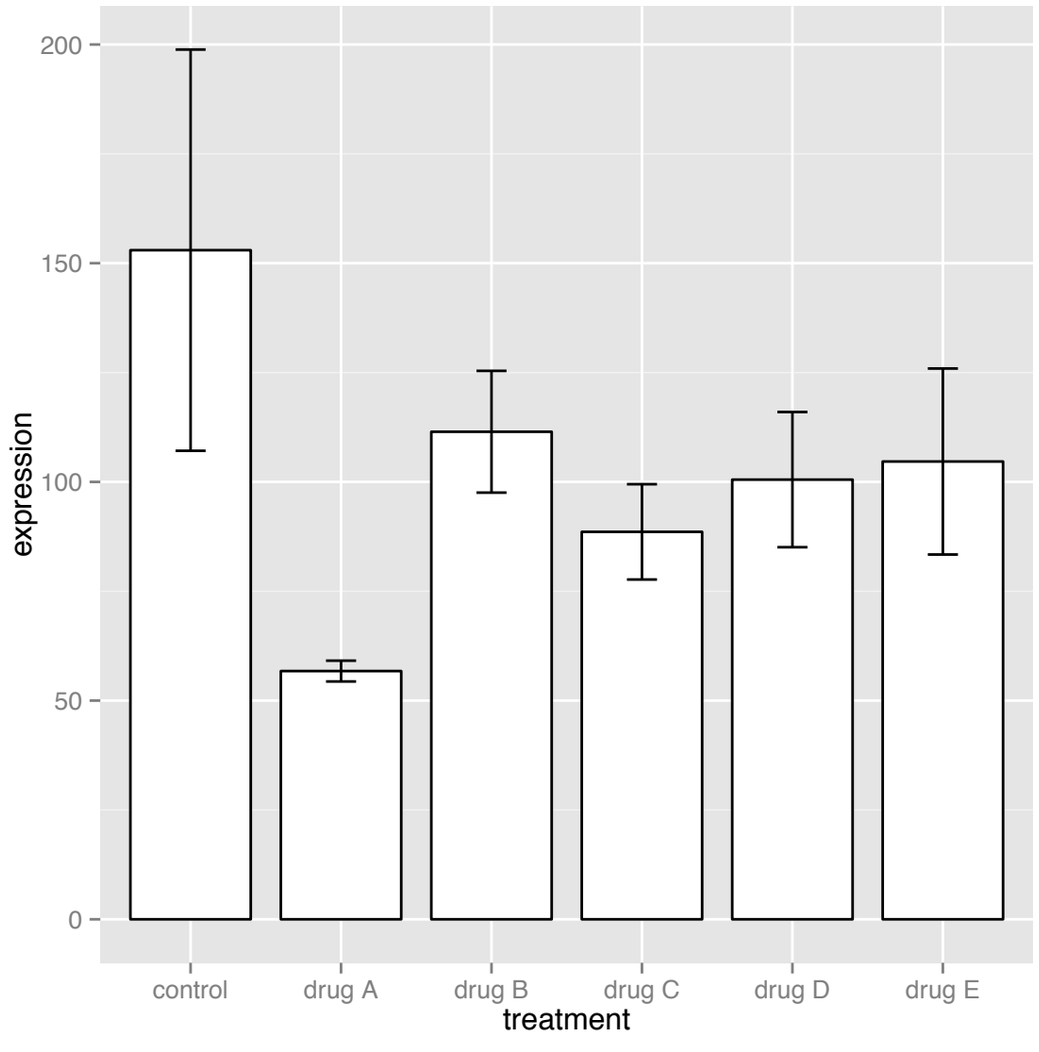
test 5 drugs on effect
on gene expression

what is the basic question?

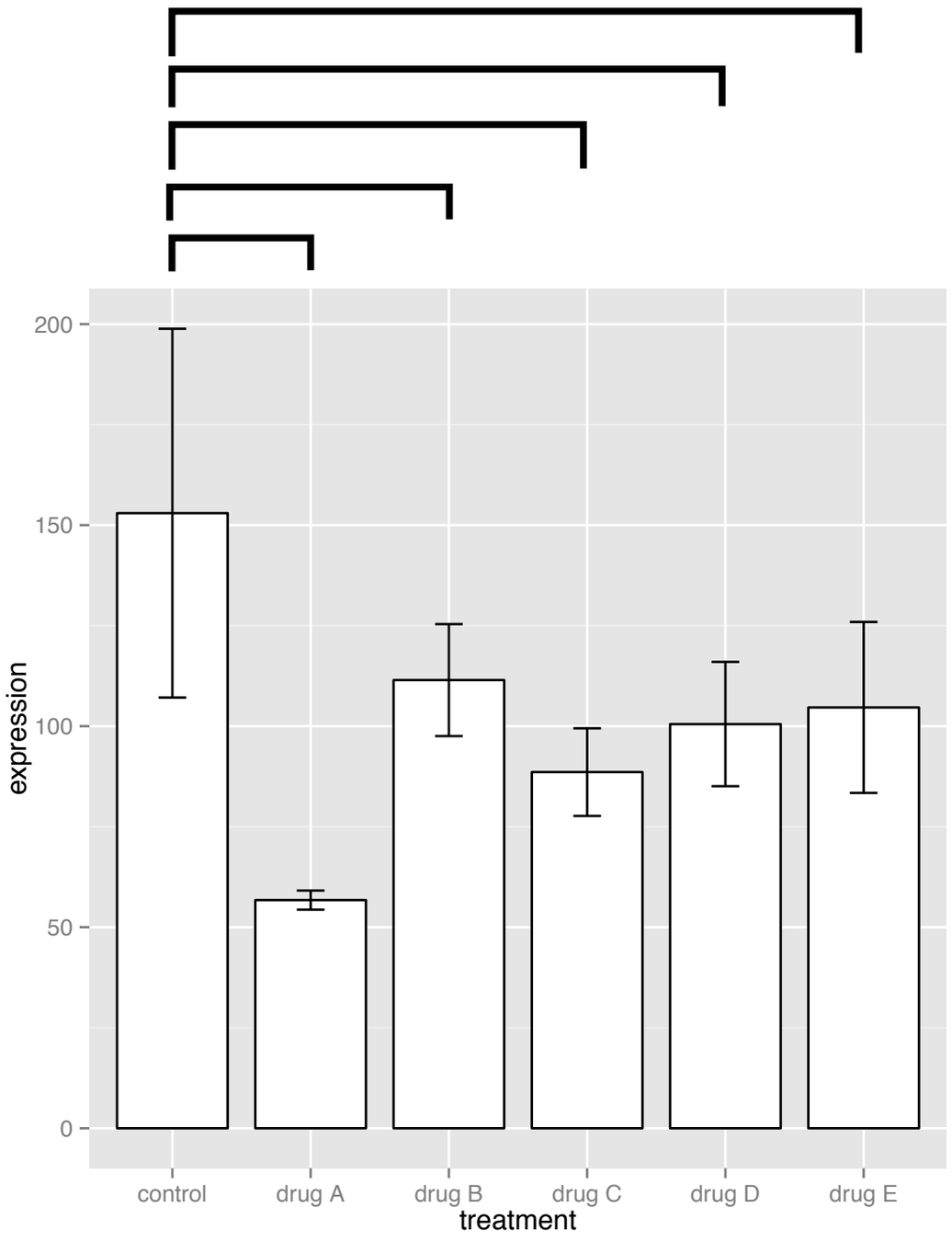


what is the basic question?

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



multiple testing

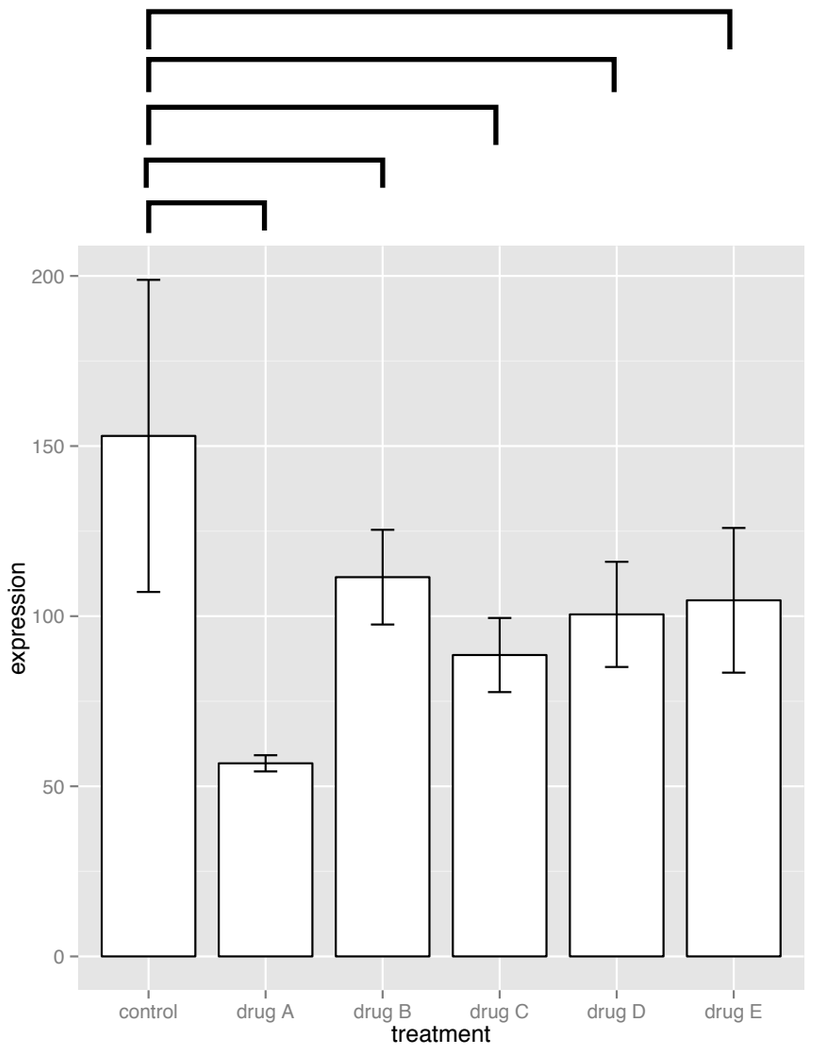


multiple testing

- inflates the type I error rate:
  error rates add up with every test conducted within an experiment
  in our case 5 t-tests each conducted at an alpha of 5% will yield an overall error rate of 25%
- if type I error should be controlled, multiple testing correction procedures have to be applied
- multiple testing typically reduces the power of the experimental setting (the more tests the lower the power)

1-way ANOVA with Dunnett's test

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



ANOVA/Dunnett requirements

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

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 ($n = 20$)	MJ ($n = 20$)	p-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
TIPI[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.

[a]State Trait Anxiety Inventory Form (Spielberger et al., 1983).

[b]Hamilton Depression Rating Scale (Hamilton, 1960).

[c]Ten-Item Personality Inventory (Gosling et al., 2003).

[d]Occasional smokers reported from 1 cigarette/week to 1 cigarette every 3 months.

confounding

Table 1. Participant demographics

	CON ($n = 20$)	MJ ($n = 20$)	p-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
TIPI[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.

[a]State Trait Anxiety Inventory Form (Spielberger et al., 1983).

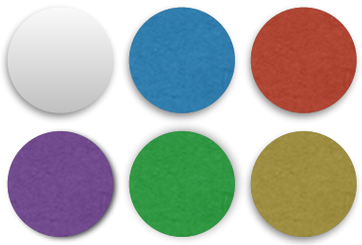
[b]Hamilton Depression Rating Scale (Hamilton, 1960).

[c]Ten-Item Personality Inventory (Gosling et al., 2003).

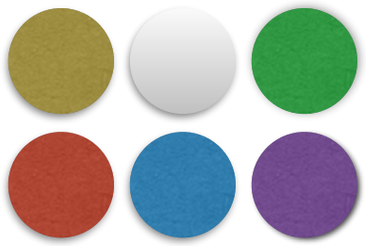
[d]Occasional smokers reported from 1 cigarette/week to 1 cigarette every 3 months.

block design, n=5, random

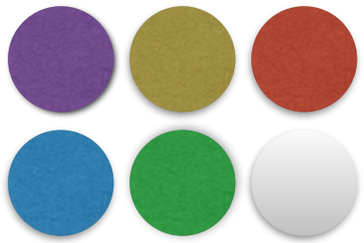
monday

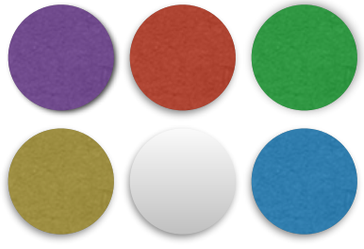


tuesday

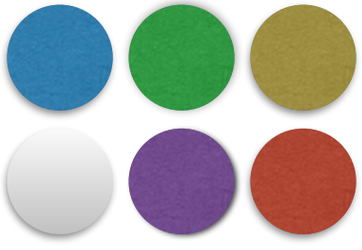


wednesday





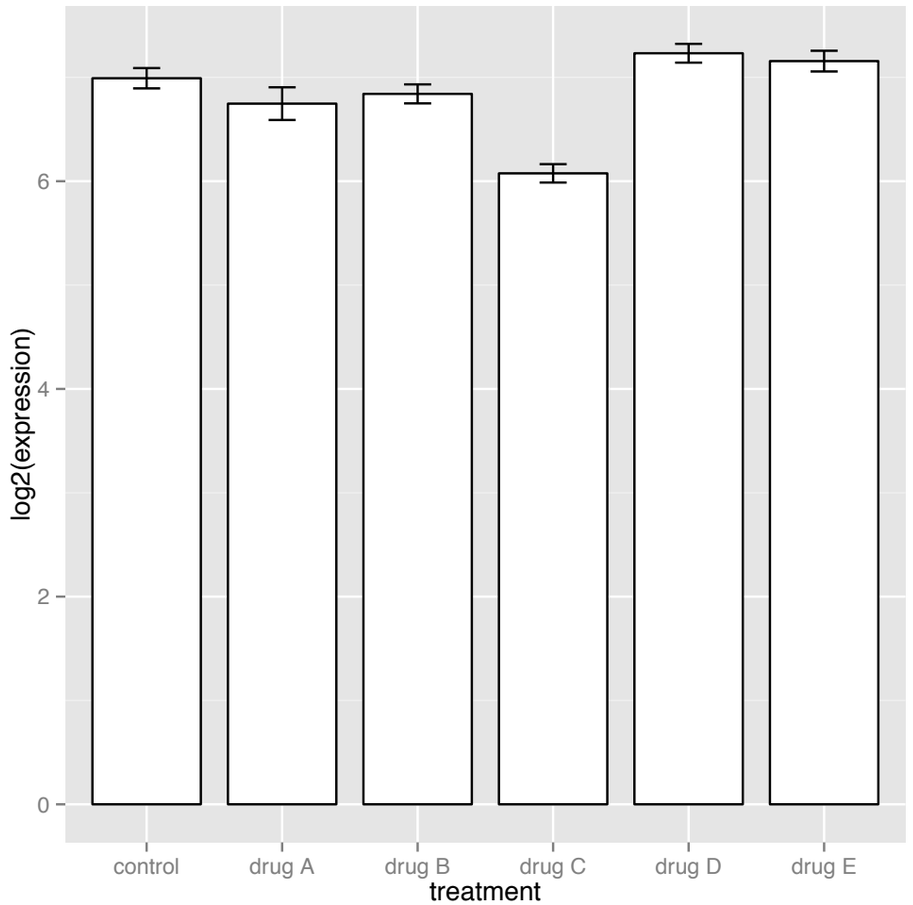
thursday



friday

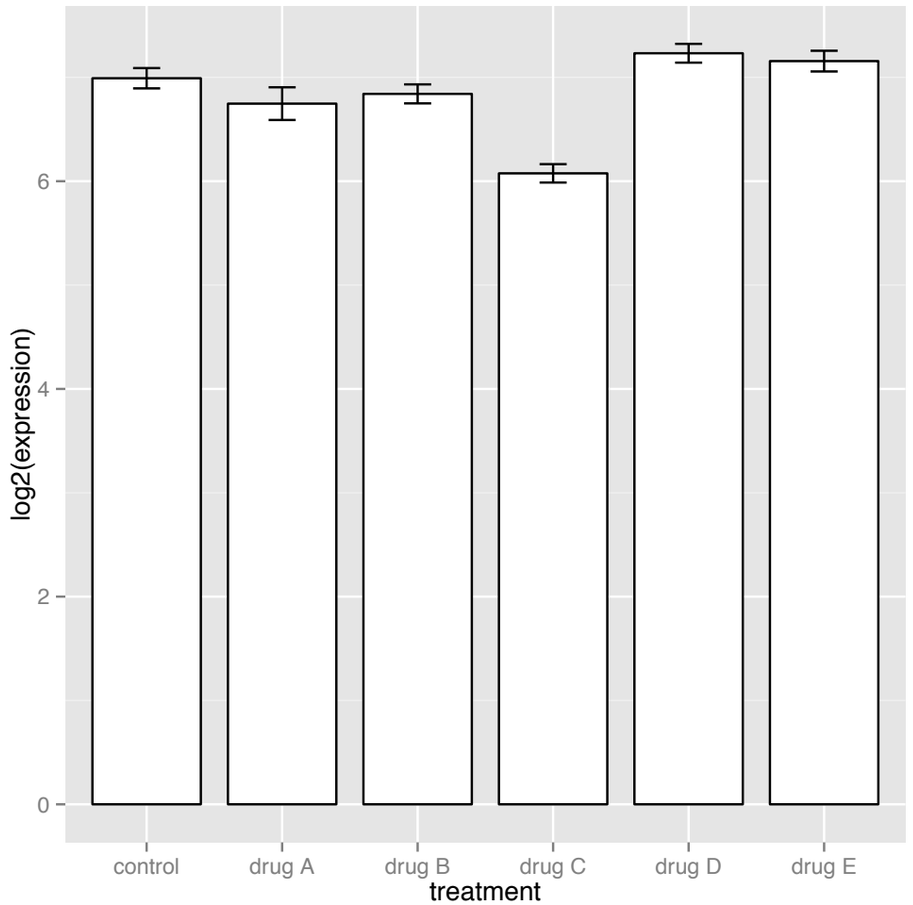
http://www.randomizer.org/form.htm

Performing the experiment



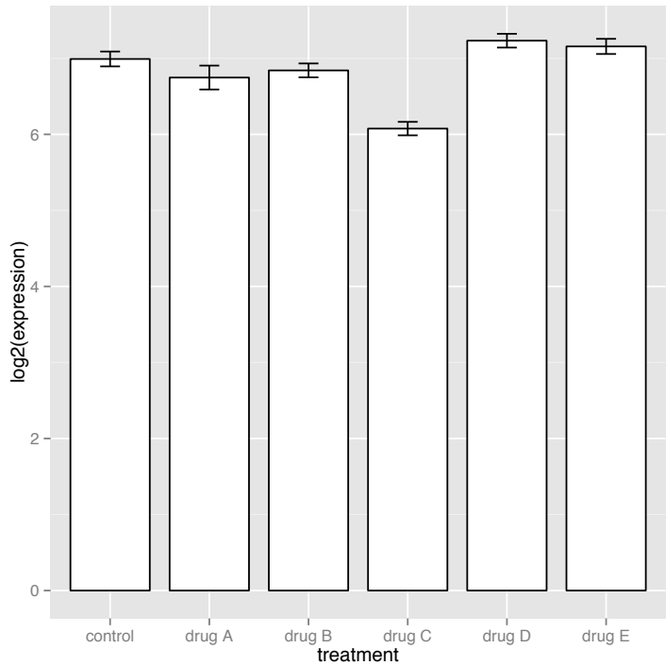
n=5,
error bars are SEM

Performing the experiment



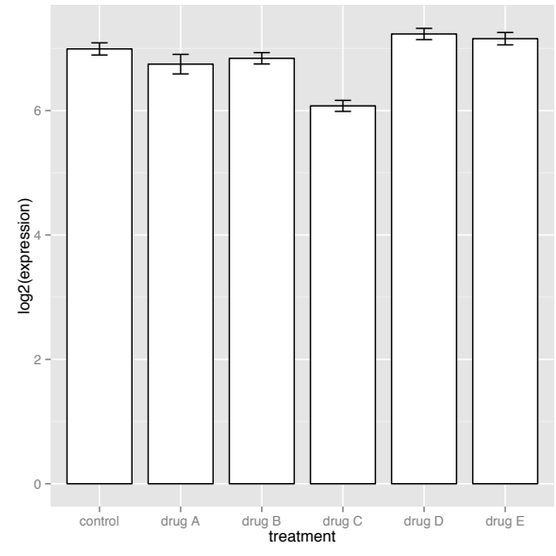
n=5,
error bars are SEM

omnibus ANOVA




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

Dunnett's test




```
     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


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

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

report - the figure

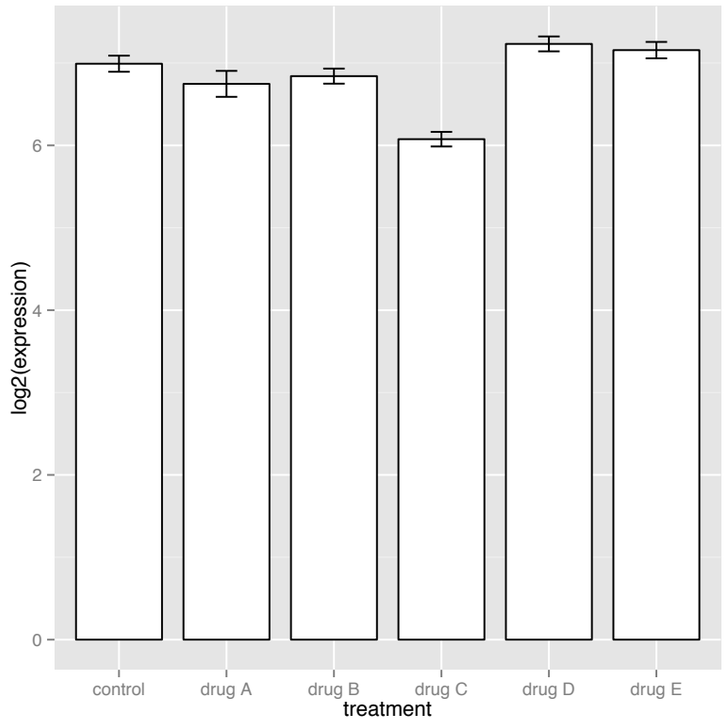


Fig.1: Drug C inhibits expression of gene x. RT-qPCR measurement of gene x transcript levels upon administration of a solvent control or 10 μM of drug A to proliferating XYZ cells for 24 hours. Error bars indicate the SEM of biological replicates (n=5).

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

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

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

statistical significance
does not equal
biological relevance

. . . and vice versa

problems of p-values

estimate	ci.low	ci.high	pval
-1.15	-2.04	-0.27	0.019
-1.23	-2.44	-0.01	0.049
-1.39	-1.89	-0.89	0.000
-0.35	-1.23	0.53	**0.367**
-0.92	-1.36	-0.48	0.001
-0.61	-1.45	0.24	**0.138**
-1.30	-1.71	-0.89	0.000
-0.41	-0.89	0.07	**0.083**
-1.04	-2.22	0.13	**0.073**
-0.60	-1.52	0.31	**0.164**
-0.85	-1.74	0.03	**0.057**
-1.03	-1.78	-0.27	0.016
-0.80	-1.43	-0.18	0.018
-0.88	-1.77	0.02	**0.055**
-1.51	-1.89	-1.13	0.000
-0.97	-1.88	-0.07	0.038
-1.10	-2.00	-0.19	0.025
-1.37	-2.03	-0.72	0.001
-1.30	-1.88	-0.72	0.001
-1.34	-2.07	-0.61	0.004
-1.21	-1.99	-0.42	0.011
-1.25	-1.53	-0.98	0.000
-0.67	-1.41	0.07	**0.068**
-1.44	-2.14	-0.74	0.003
-1.30	-2.18	-0.41	0.010
-1.14	-1.61	-0.67	0.001
-0.94	-1.86	-0.02	**0.047**
-1.41	-2.14	-0.69	0.003
-0.80	-1.31	-0.29	0.007
-0.65	-1.69	0.38	0.179

problems of p-values

- p-values are highly unreliable (irreproducible) even at large n!
- p-values do not reveal the underlying effect size
- confidence intervals are better descriptors of the robustness and extend of effects

a p-value is a p-value

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

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

Suggestion to authors

- Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article
- Authors must list all variables collected in a study
- Authors must report all experimental conditions, including failed manipulations

the Jens Förster case

"if the data did not confirm the hypothesis, I talked to people in the lab about what needs to be done next, which would typically involve brainstorming about what needs to be changed, implementing the changes, preparing the new study and re-running it"

research types

- **exploratory research**
  - hypothesis generating
  - no/little prior information on effects, frequently many endpoints measured
  - often not complying with elementary rules of sampling and experimental layout (e.g. sequential sampling)
  - statistical testing will yield highly problematic results (low power, high error rate), potentially irreproducible
- **confirmatory research**
  - performed to confirm hypotheses
  - solid prior knowledge on effects
  - involves prior power analysis, thoughtful experimental layout
  - generates more reliable statistical test results, potentially reproducible

a pragmatic solution

- in basic exploratory research - “discovering something new” - we cannot generate high confidence results that are likely to be reproducible. (N is low, statistical power is poor)
- representative/unbiased sampling is fundamentally important
- instead of reporting p-values we should mainly focus on reporting the effect size (or, if inference is desired, confidence intervals).
- multiple testing correction and any other complex statistical treatments/tests should be simply omitted.
- Ask simple questions and perform simple tests.

a pragmatic solution

- if one wishes to obtain a higher certainty rules for confirmatory research apply
- prerequisite is prior information given e.g. by pioneering experiments for the estimation of the effect size
- ideally a more complex experimental setting should be reduced to a simple 2-level comparative study
- thorough experimental design and an a-priori-power analysis has to be performed

a pragmatic solution for interpretation

- when evaluating exploratory research results, which are probably the vast majority of results in basic life science research, we have to keep their limitations in mind (i.e. the p-values are pretty meaningless).
- But still: it is the data that matters, not the story.

Experimental design

- Aim:
  - Generalisation, Inference, Induction
  - Elimination of systematic errors and non-biological variances (noise)
    - Estimation of the ‘biological effect’
- Design has to be set up before data collection
- Important means:
  - Manipulative study (comparing untreated versus treated)
  - Sample independence, representativity,
  - randomization (increases accuracy), replication (increases precision)
  - blinding

TABLE 1. Potential sources of confusion in an experiment and means for minimizing their effect.

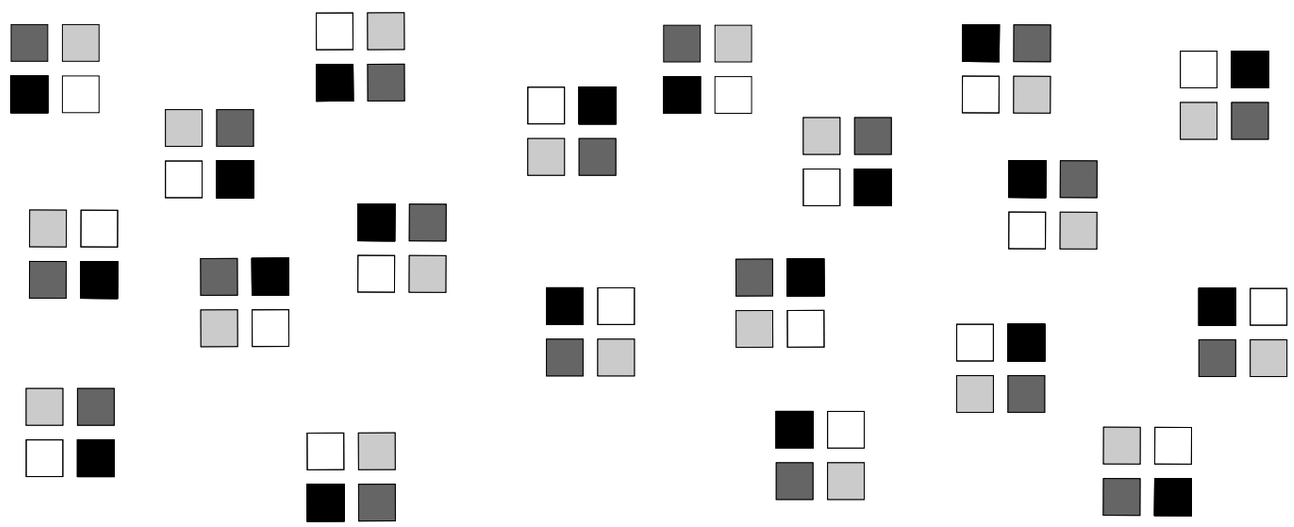
Source of confusion	Features of an experimental design that reduce or eliminate confusion
1. Temporal change	Control treatments
2. Procedure effects	Control treatments
3. Experimenter bias	Randomized assignment of experimental units to treatments Randomization in conduct of other procedures “Blind” procedures*
4. Experimenter-generated variability (random error)	Replication of treatments
5. Initial or inherent variability among experimental units	Replication of treatments Interspersion of treatments Concomitant observations
6. Nondemonic intrusion	Replication of treatments Interspersion of treatments
7. Demonic intrusion	Eternal vigilance, exorcism, human sacrifices, etc.

* Usually employed only where measurement involves a large subjective element.

† Nondemonic intrusion is defined as the impingement of chance events on an experiment in progress.

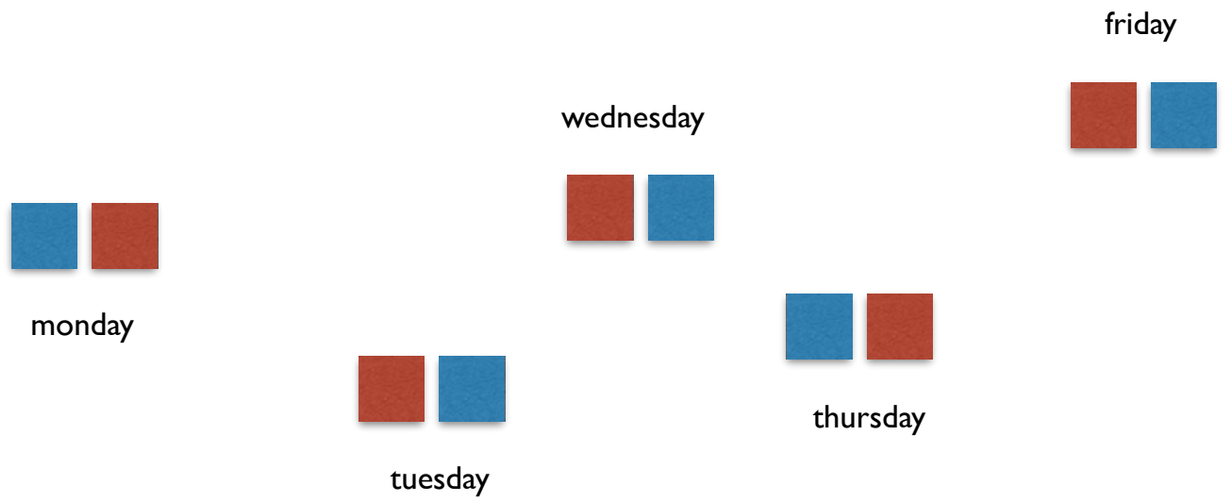
Hurlbert SH. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **54(2)**: 187–211.

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 noise 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

Documentation

- DFG: “Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, zehn Jahre lang aufbewahrt werden.”
- Always keep the raw data (measurement results, unprocessed images). Ideally the raw data should be part of publications.
- All experimental details (including computational analysis codes) have to be documented and ideally made available in publications.
- Raw data and experimental details should be disclosed among research collaborators.

Towards reproducible research

- Familiarise yourself with the basic concepts of statistics and experimental design.
- Try to test simple hypotheses.
- Sample in an unbiased way.
- Keep the raw data and make it available to others.
- Report confidence intervals (of effects) and N.
- Be the most critical judge over your own data.
- Don’t trust p-values. Not at all.

useful links

- http://udel.edu/~mcdonald/statintro.html
- http://www.randomizer.org/form.htm
- http://www.statisticalsolutions.net/pss_calc.php
- http://www.wormbook.org/chapters/www_statisticalanalysis/statisticalanalysis.html
- www.statisticsdonewrong.com