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

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 $2 \times 2$

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>$0.01$</td>
<td>$0.99$</td>
</tr>
<tr>
<td>Negative</td>
<td>$0.99$</td>
<td>$0.01$</td>
</tr>
</tbody>
</table>

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


Experimenter’s contribution to reproducible research

- Experimental Design
- Statistics
- Documentation
- Interpretation
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

<table>
<thead>
<tr>
<th></th>
<th>control</th>
<th>drug A</th>
</tr>
</thead>
<tbody>
<tr>
<td>198.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>106.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>153.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>56.10</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>59.39</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>54.77</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>191.08</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>170.11</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>197.10</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>112.23</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>134.80</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>130.17</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>120.60</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>173.16</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>120.58</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>64.20</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>87.30</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>86.11</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>105.76</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>108.74</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>84.84</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>43.52</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>49.50</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>49.58</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>144.15</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>117.92</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>148.93</td>
<td>control</td>
<td></td>
</tr>
<tr>
<td>70.14</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>52.66</td>
<td>drug A</td>
<td></td>
</tr>
<tr>
<td>67.89</td>
<td>drug A</td>
<td></td>
</tr>
</tbody>
</table>

### Unpaired t test with equal SD

<table>
<thead>
<tr>
<th>Table Analyzed</th>
<th>Drug</th>
<th>P value</th>
<th>P value summary</th>
<th>Significantly different? (P &lt; 0.05)</th>
<th>One- or two-tailed P value?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>&lt; 0.0001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>1.549, 14, 14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>0.4229</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Significantly different? (P < 0.05)

- Yes
- No

### One- or two-tailed P value?

- Two-tailed
- One-tailed

### t, df

- t=5.592 df=28

### Confidence interval

- 93.23 to 43.24

### R squared

- 0.5276

### Effect size

- 43.24

### Mean ± SEM

- Column A: 142.8 ± 9.512
- Column B: 74.56 ± 7.642

### Difference between means

- -68.24 ± 12.20

### P value

- 0.029

### Mean ± SEM of column

- 142.8 ± 9.512, n=15
- 74.56 ± 7.642, n=15

### F test to compare variances

- 1.549, 14, 14

### P value summary

- ns
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 [...]
the manuscript

**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 to Reference [1]. Statistical analysis was done in GraphPad 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
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
Replicability

- 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

Reproducibility

- Reproduction using different reagents/tools but the same protocol by a different person outside the lab
- Reproduction just based on text description

POPULATION
Replicability

Reproduction of the original results using the same protocol/reagents/tools by the same person.

Reproducibility

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

Reproduction just based on text description.

Several methodological factors 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
Inference

Howto
starting from the population

gene X expression (population)

expression
mean: 135.917
starting from the population

gene X expression (population)

log(expression)
mean: 7
population

log(expression)
mean: 7

expt 1
mean: 7.135

expt 2
mean: 6.274

expt 3
mean: 7.172
population

log(expression)

Frequency

0e+00 4e+05 8e+05

n=3

mean: 7.238

log(expression)

n=10

mean: 6.908

log(expression)

n=30

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.

\[ \text{SEM} = \frac{\text{standard deviation}}{\sqrt{n}} \]

- 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:

- **Mean**: 7
- **Log(expression)**

<table>
<thead>
<tr>
<th>Frequency</th>
<th>0e+00</th>
<th>4e+05</th>
<th>8e+05</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**n=3**

- **Log(expression)**
  - **Mean**: 6.698

**n=10**

- **Log(expression)**
  - **Mean**: 6.974

**n=30**

- **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%.

- $(1−\alpha)$-confidence interval: An estimated interval which contains the „true value“ of a quantity with a probability of $(1−\alpha)$.
  $1−\alpha = \text{confidence level}$, $\alpha = \text{error probability}$
The histograms show the frequency distribution of log(expression) for the population and three different sample sizes (n=3, n=10, n=30). The population distribution is centered around log(7).

For n=3:
- Mean: 7.28

For n=10:
- Mean: 6.671

For n=30:
- 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?
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
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

<table>
<thead>
<tr>
<th></th>
<th>replicate 1</th>
<th>replicate 2</th>
<th>replicate 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>234</td>
<td>220</td>
<td>229</td>
</tr>
<tr>
<td>Tuesday</td>
<td>269</td>
<td>967</td>
<td>275</td>
</tr>
<tr>
<td>Wednesday</td>
<td>254</td>
<td>249</td>
<td>246</td>
</tr>
</tbody>
</table>

Example taken from "Intuitive Biostatistics," H. Motulsky
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

- 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

**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.
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

The effect

mean difference = \text{mean(treatment)} - \text{mean(control)} = -1 \ (0.5 \text{ in linear space}) = \text{THE EFFECT}
drug effect on gene X (population)

log2(expression gene X)

 ctr

 drug

 Density

0.0 0.2 0.4 0.6 0.8

5 6 7 8 9 10

n=3

log2(expression) mean: -1.392

n=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)
null distribution

log2(expression)
mean: -0.46

TEST

value

p-value

significant!!??

n=4

control

drug

4 5 6 7 8 9 10

mean: -0.46

4 5 6 7 8 9 10
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 = 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

but there was an effect!
### Error of Statistical Tests

#### Distribution of the Test Statistic

- **Distribution of the test statistic under the null hypothesis**
- **Distribution of the test statistic under the alternative hypothesis**

#### Table: Hypothesis Testing Outcomes

<table>
<thead>
<tr>
<th></th>
<th>Accept null hypothesis</th>
<th>Reject null hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Null Hypothesis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is TRUE</td>
<td>Correct decision</td>
<td>Type I Error</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“False Positive”</td>
</tr>
<tr>
<td><strong>Alternative Hypothesis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is TRUE</td>
<td>Type II Error</td>
<td>Correct decision</td>
</tr>
<tr>
<td></td>
<td>“False Negative”</td>
<td></td>
</tr>
</tbody>
</table>
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 ($\beta$). Therefore power is equal to $1 - \beta$, which is also known as the sensitivity.
power analysis

- sample size $N$
- effect size $d = \frac{\bar{x}_1 - \bar{x}_2}{s}$
- $\alpha$, significance level (0.05)
- power, $1-\beta$ (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

\( \text{power} = \)

alternative = two.sided

NOTE: n is number in *each* group
Two-sample t test power calculation

\[ n = 4 \]
\[ d = 2 \]
\[ \text{sig.level} = 0.05 \]
\[ \text{power} = 0.6568759 \]
\[ \text{alternative} = \text{two.sided} \]

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

underpowered!
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 $2 \times 2$
power analysis - obtaining N

Two-sample t test power calculation

\[
\begin{align*}
n &= \\
d &= 2 \\
sig.level &= 0.05 \\
power &= 0.8 \\
\text{alternative} &= \text{two.sided}
\end{align*}
\]

NOTE: n is number in *each* group
power analysis - obtaining N

Two-sample t test power calculation

\[ n = 5.090002 \]
\[ d = 2 \]
\[ \text{sig.level} = 0.05 \]
\[ \text{power} = 0.8 \]
\[ \text{alternative} = \text{two.sided} \]

**NOTE:** n is number in *each* group
overpower!

Two-sample t test power calculation

\[ n = 10000 \]
\[ d = \]
\[ \text{sig.level} = 0.05 \]
\[ \text{power} = 0.8 \]
\[ \text{alternative} = \text{two.sided} \]

NOTE: \( n \) is number in *each* group
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).
I-way ANOVA with Dunnett’s test

- omnibus I-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

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

Even casually smoking marijuana can change your brain, study says
Table 1. Participant demographics

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

All values are expressed in means and SDs. CON, controls; MJ, marijuana users.
<sup>+</sup>State Trait Anxiety Inventory Form (Spielberger et al., 1983).
<sup>+</sup>Hamilton Depression Rating Scale (Hamilton, 1960).
<sup>+</sup>Ten-Item Personality Inventory (Gosling et al., 2003).
<sup>+</sup>Occasional smokers reported from 1 cigarette/week to 1 cigarette every 3 months.
Table 1. Participant demographics

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

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

<sup>a</sup>State Trait Anxiety Inventory Form (Spielberger et al., 1983).
<sup>b</sup>Hamilton Depression Rating Scale (Hamilton, 1960).
<sup>c</sup>Ten-Item Personality Inventory (Gosling et al., 2003).
<sup>d</sup>Occasional smokers reported from 1 cigarette/week to 1 cigarette every 3 months.
block design, n=5, random

monday

wednesday

tuesday

thursday

friday

http://www.randomizer.org/form.htm
Performing the experiment

\[ \text{log}_2(\text{expression}) \]

\( n=5, \)
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:

| Drug     | Estimate | Std. Error | t value | Pr(>|t|) |
|----------|----------|------------|---------|----------|
| drug A   | -0.2446  | 0.1513     | -1.617  | 0.352    |
| drug B   | -0.1505  | 0.1513     | -0.995  | 0.777    |
| drug C   | -0.9158  | 0.1513     | -6.053  | <1e-04 *** |
| drug D   | 0.2406   | 0.1513     | 1.590   | 0.368    |
| drug E   | 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)
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)) […]

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.

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

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


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

* 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.

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