Lecture 5

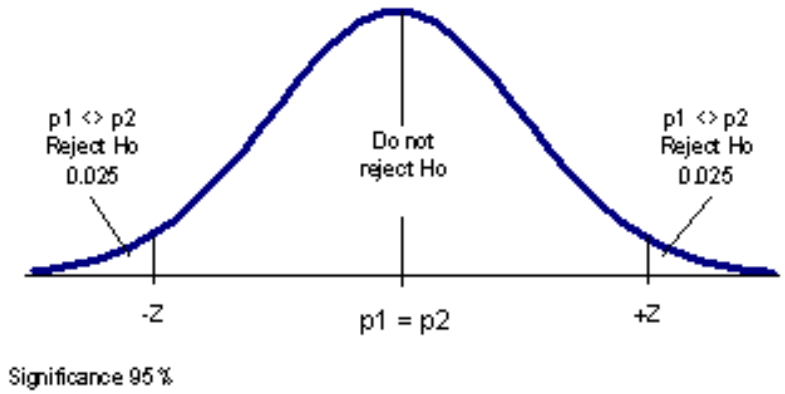
Continuing on significance tests

Criticisms of the NHST

Publication bias

Research planning

Theophanis Tsandilas



p-value

The *p*-value is the probability of obtaining a statistic *as extreme or more extreme than the one observed* if the null hypothesis was true.

Normal distributions

If the sampling distribution of the statistic is normal, we will use the standard normal distribution z to derive the p value

Example

An experiment studies the IQ scores of people lacking enough sleep.

$H_0: \mu = 100$ and $H_1: \mu < 100$ **(one-sided)**

or

$H_0: \mu = 100$ and $H_1: \mu \neq 100$ **(two-sided)**

Example

Results from a sample of 15 participants are as follows:

90, 91, 93, 100, 101, 88, 98, 100, 87, 83, 97, 105, 99, 91, 81

The mean IQ score of the above sample is **M = 93.6.**

Is this value **statistically significantly different** than 100?

Creating the test statistic

We assume that the population standard deviation is known and equal to SD = 15. Then, the standard error of the mean is:

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{15}} = 3.88$$

Creating the test statistic

We assume that the population standard deviation is known and equal to SD = 15. Then, the standard error of the mean is:

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{15}} = 3.88$$

The **test statistic** tests the **standardized** difference between the observed mean $\hat{\mu} = 93.6$ and $\mu_0 = 100$

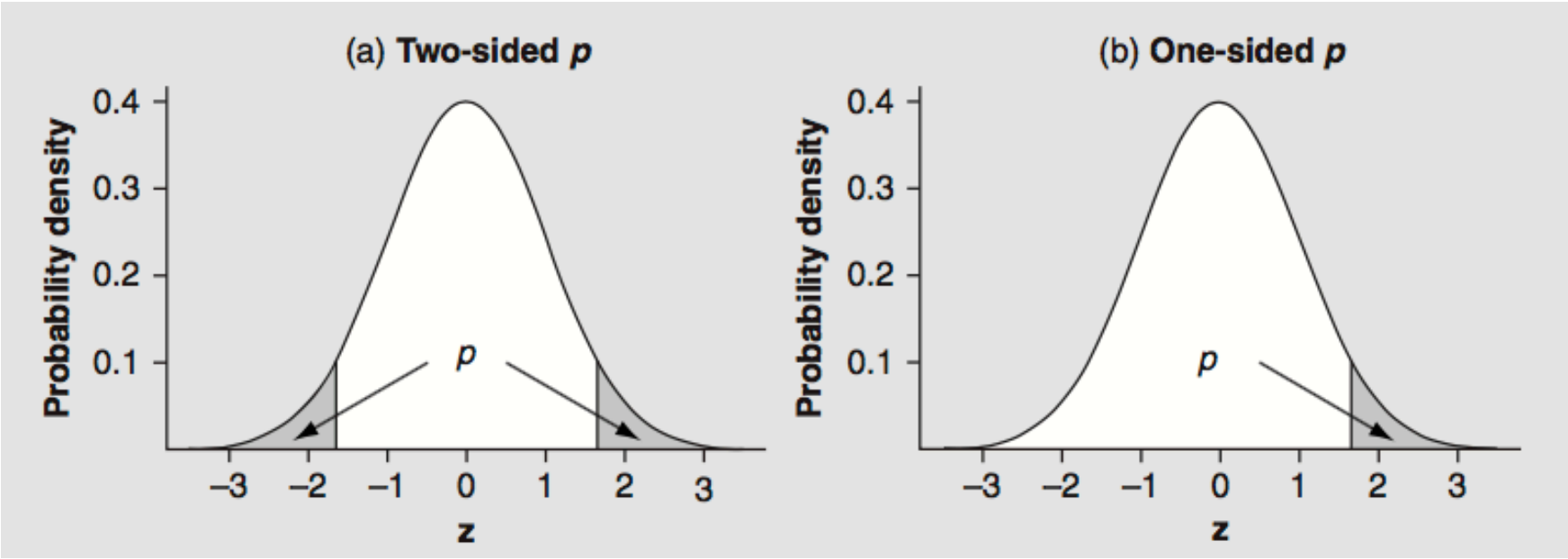
$$z = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} = \frac{93.6 - 100}{3.88} = -1.65$$

The *p* value is the probability of getting a z statistic as or more extreme than this value (given that H_0 is true)

Calculating the p value

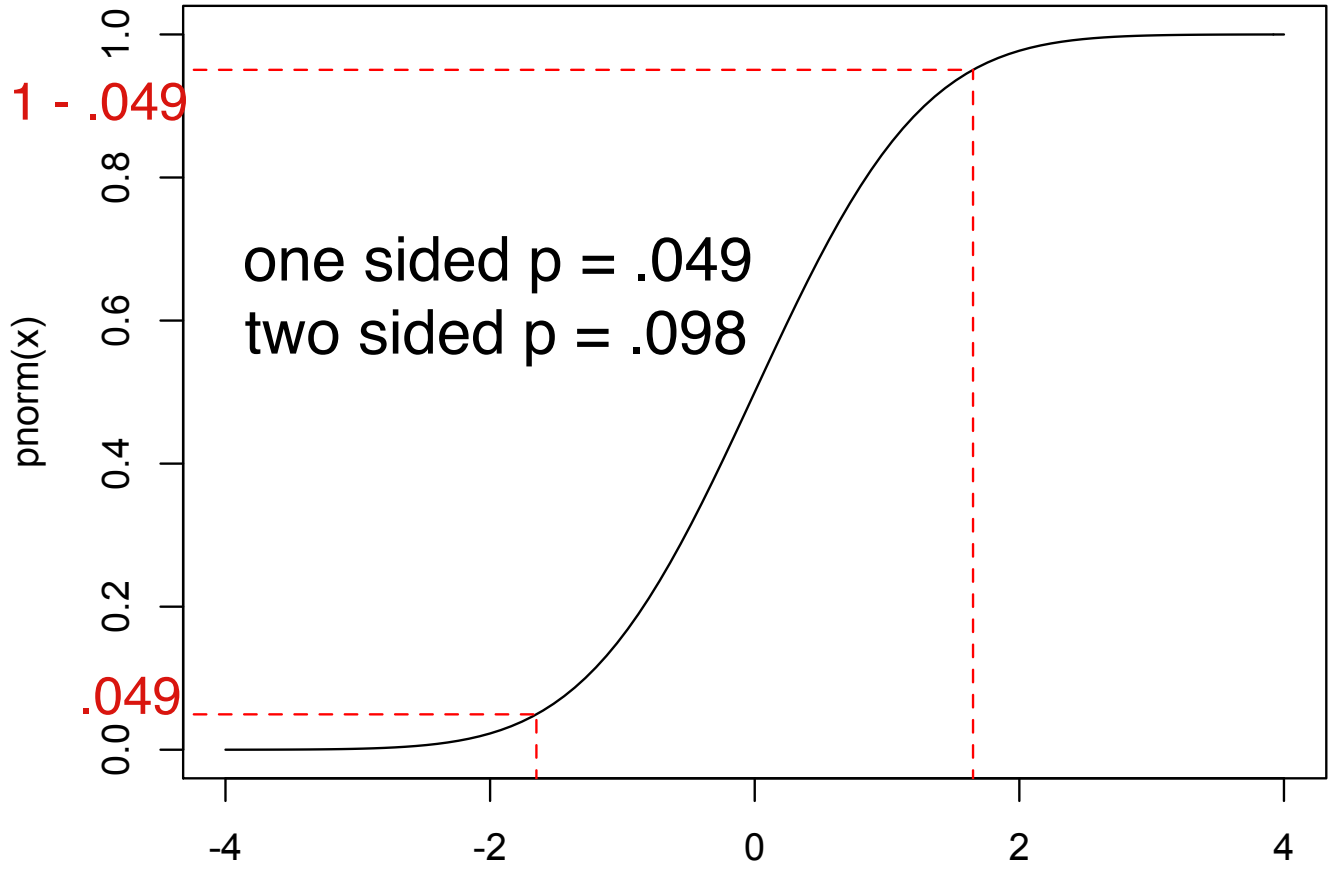
$$z = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} = \frac{93.6 - 100}{3.88} = -1.65$$

The *p* value is the probability of getting a z statistic as or more extreme than this value (given that H_0 is true)



Calculating the p value

To calculate the area in the distribution, we will work with the cumulative density probability function (cdf).



R code

```r
m0 <- 100 # mean for null hypothesis
sd <- 15 # We assume that the population sd is known

# These are the observed IQ scores
scores <- c(90, 91, 93, 100, 101, 88, 98, 100, 87, 83, 97, 105,
99, 91, 81)
m <- mean(scores)

# I calculate the standard error of the mean
se <- sd / sqrt(length(scores))

# I calculate the z statistic, i.e., the standardized mean
difference
z = (m - m0) / se

# This is the one side p value
pvalue <- pnorm(z)

cat("one sided p-value =", pvalue, "\n")
cat("two sided p-value =", 2*pvalue, "\n")
```

Example: conclusions

If we assume that the standard deviation of the population is known (SD = 15):

We reject the null hypothesis ($\alpha = .05$) if we apply a one-sided significance test ($p = .049$).

But we cannot reject it if we apply a two-sided significance test ($p = .092$)

One-sided vs. two sided

“A common misunderstanding of directional testing is to believe that it should be employed whenever a researcher predicts an effect in a particular direction.

This is almost always bad practice. *A one-sided test should be employed only if the direction of an effect is already known or if any outcome in the non-predicted direction would be ignored.*

The crucial question is not whether you think an effect lies in a particular direction, but whether you are willing to declare an effect in the wrong direction non-significant (no matter how interesting or how important it is). The answer to such a question is usually ‘no’ and for this reason ***one-sided tests should typically be avoided.****“*

from Baguley

t tests

If the population variance is unknown, it is better to work with the *t* distribution.

One sample *t*

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

$$t = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}} = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

Where $\hat{\sigma}$ is the unbiased estimate of the population standard deviation.

And t follows a *t* distribution with *v* = *n - 1* degrees of freedom.

Back to our example

Results from a sample of 15 participants are as follows:

90, 91, 93, 100, 101, 88, 98, 100, 87, 83, 97, 105, 99, 91, 81

The mean IQ score of the above sample is **M = 93.6.** Is this value statistically significantly different than 100?

Assume now that we don't have any prior knowledge of the population standard deviation.

R code

```
m0 <- 100 # mean for null hypothesis

# These are the observed IQ scores
scores <- c(90, 91, 93, 100, 101, 88, 98, 100, 87, 83, 97, 105,
99, 91, 81)
n <- length(scores)
m <- mean(scores)

# I estimate the standard error of the mean
se <- sd(scores) / sqrt(n)

# I calculate the t statistic, i.e., the standardized mean
difference
t = (m - m0) / se

# This is the two-sided p value, calculated from the cumulative
density function of the t distribution
pvalue <- 2*pt(t, n - 1)
cat("two sided p-value =", pvalue, "\n")
```

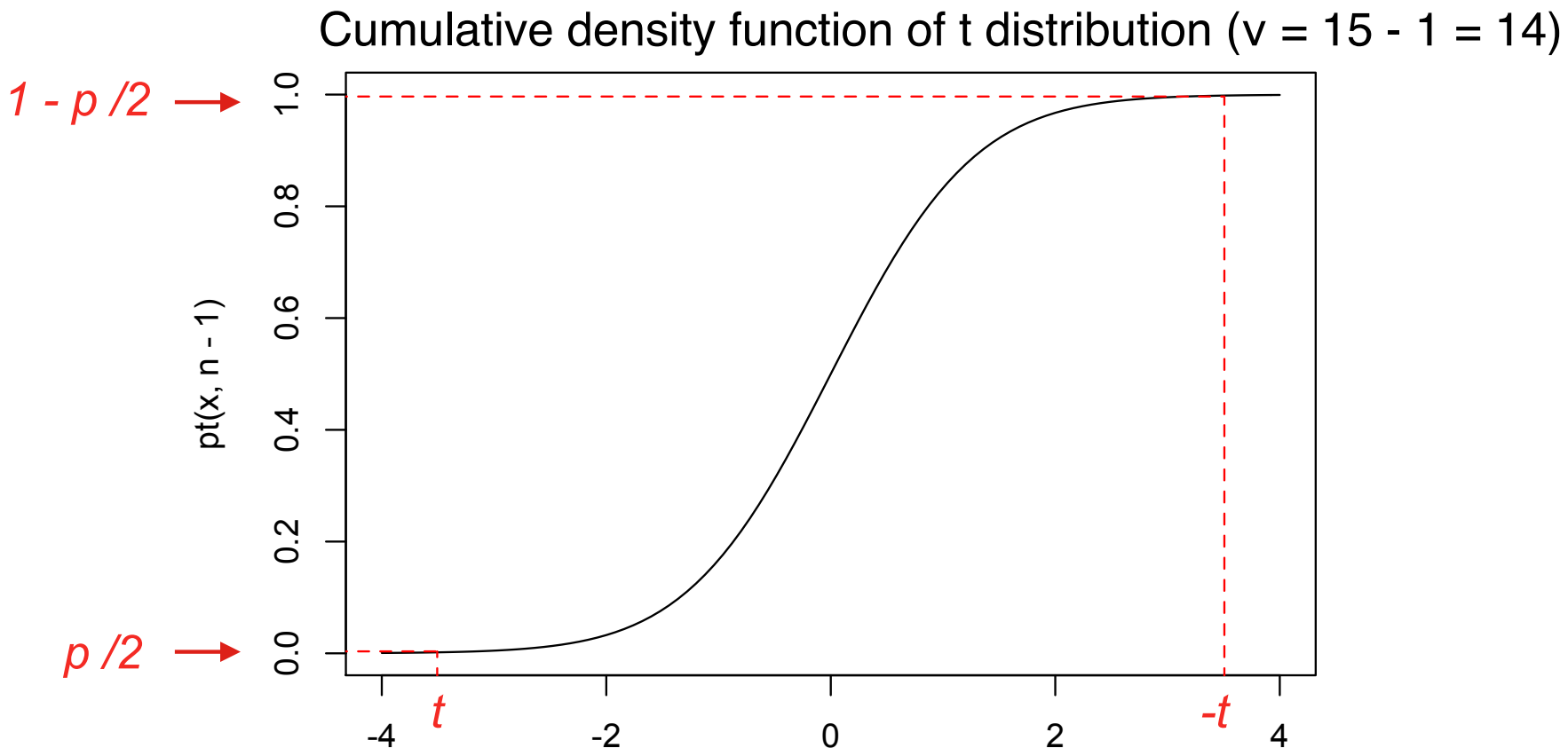
R code

Alternatively, you can use the t-test function of R:

```
> t.test(scores - 100)

   One Sample t-test

data:  scores - 100
t = -3.5064, df = 14, p-value = 0.00349
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -10.314708  -2.485292
sample estimates:
mean of x
     -6.4
```

Calculating the p value (two-sided)

$$t = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}} = -3.5064$$



Example: summarizing the result

$t(14) = -3.51$, $p = .003$

The researchers reject the null hypothesis. They found that the lack of sleep leads to a mean IQ score that is statistically significantly lower than 100.

Independent *t* test

NHST of a difference between the means of two independent normal samples.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_{\hat{\mu}_1 - \hat{\mu}_2}} \sim t(\nu)$$

If we assume equal variances, then:

$$\nu = (n_1 + n_2) - 2 \quad \text{and} \quad \hat{\sigma}_{\hat{\mu}_1 - \hat{\mu}_2} = \hat{\sigma}_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Independent *t* test

NHST of a difference between the means of two independent normal samples.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_{\hat{\mu}_1 - \hat{\mu}_2}} \sim t(\nu)$$

But if variances are unequal, the Welch-Satterthwaite correction should be used. Please, check the slides of Lecture 3 for more details.

Example

A research team is interested in comparing the performance in an IQ test between two groups: adults who have (G1) and adults who have not (G2) completed any graduate studies.

From each group, they test 15 participants.

S_{G1}: 102, 104, 103, 106, 95, 108, 101, 108, 113, 96, 112, 106, 105, 109, 105

S_{G2}: 96, 108, 90, 104, 97, 103, 95, 102, 93, 107, 101, 88, 99, 104, 97

R code

```
> scores1 <- c(102, 104, 103, 106, 95, 108, 101, 108, 113, 96, 112, 106, 105, 109, 105)
> scores2 <- c(96, 108, 90, 104, 97, 103, 95, 102, 93, 107, 101, 88, 99, 104, 97)
> t.test(scores1, scores2)

   Welch Two Sample t-test

data:  scores1 and scores2
t = 2.9411, df = 27.316, p-value = 0.006589
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.796244 10.070423
sample estimates:
mean of x mean of y
104.86667  98.93333
```

Corrected for unequal variances by default

Example: summarizing the results

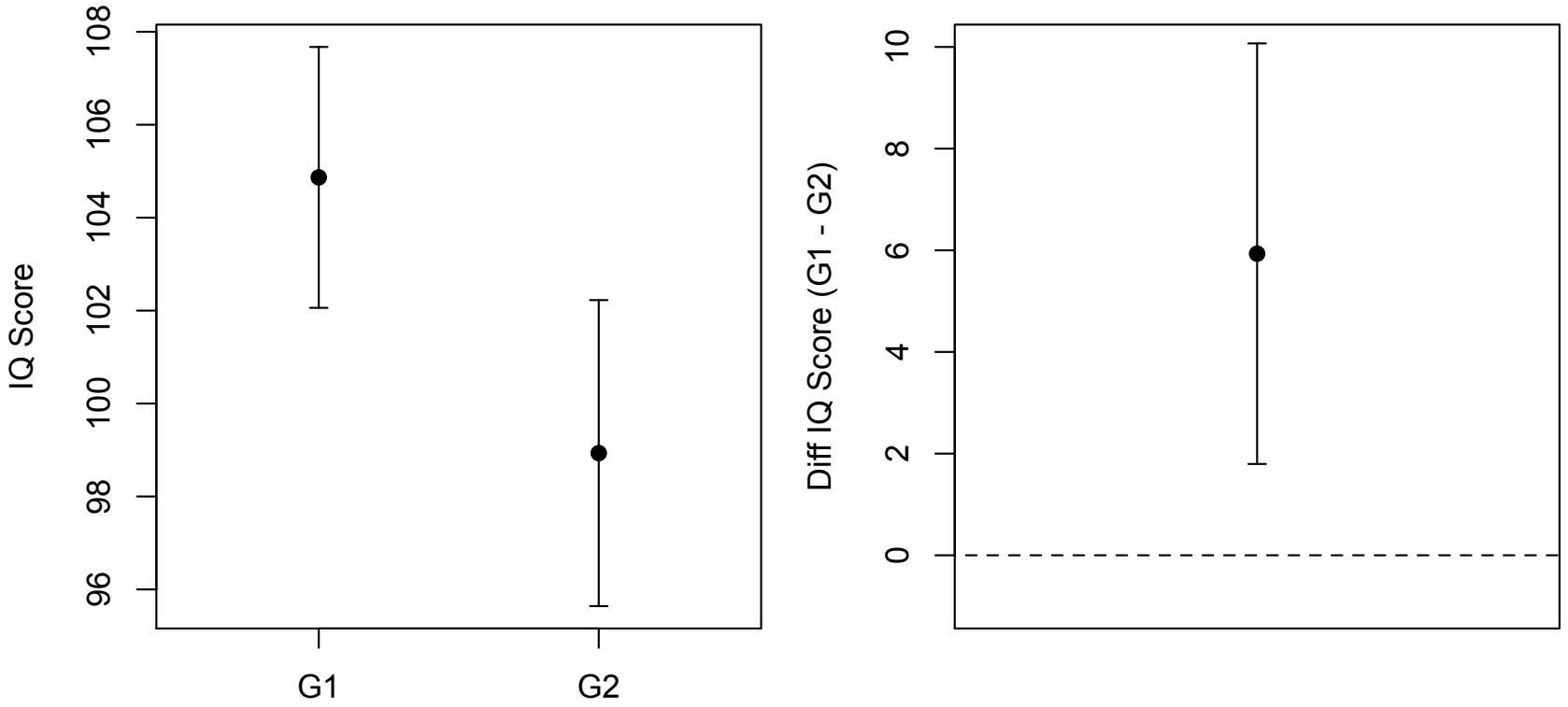
$t(27.32) = 2.94$, $p = .007$

The researchers reject the null hypothesis ($\alpha = .05$).

They conclude that the mean IQ score of adults who have completed graduate studies (G1) is statistically significantly higher than the IQ score of adults who have not completed any graduate studies (G2).

Example: summarizing the results

The 95% confidence intervals for this same example.



Paired observations

A research team is interested in comparing the performance in an IQ test before and after attending a Geometry course:

10 participants are tested before and after the completion of the course.

S_{Before}: 102, 94, 90, 104, 95, 100, 101, 96, 100, 96

S_{After}: 106, 108, 93, 103, 100, 100, 105, 98, 103, 96

Paired *t* test

The solution is again trivial. We calculate the difference in performance for each participant.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Before	102	94	90	104	95	100	101	96	100	96
After	106	108	93	103	100	100	105	98	103	96
Diff.	4	14	3	-1	5	0	4	2	3	0

and then perform the one sample *t* test on these differences.

R code

```
> scores.before <- c(102,  94, 90, 104, 95, 100, 101, 96, 100, 96)
> scores.after <- c(106, 108, 93, 103, 100, 100, 105, 98, 103, 96)
> t.test(scores.after - scores.before)

   One Sample t-test

data:  scores.after - scores.before
t = 2.5468, df = 9, p-value = 0.03136
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.3800225 6.4199775
sample estimates:
mean of x
      3.4

> t.test(scores.after, scores.before, paired = TRUE)

   Paired t-test

data:  scores.after and scores.before
t = 2.5468, df = 9, p-value = 0.03136
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3800225 6.4199775
sample estimates:
mean of the differences
                    3.4
```

Lognormal distributions

We follow the same approach as for confidence intervals:

We simply apply the t tests on the logarithms of the original values and reject the null hypothesis if $p < \alpha$.

(but beware of the interpretation of your results)

Reporting results of statistical tests

APA's (American Psychology Association) style:

Report the exact *p* value within the test
(unless the *p* value is less than .001)

$t(7) = 1.92, p = .022$

$t(54) = 5.43, p < .001$

$t(19) = -0.75, p = .460$

Two decimal points

In other research disciplines, lower thresholds may be used.

See: https://www.statisticshowto.com/probability-and-statistics/reporting-statistics-apa-style/

Reporting results of statistical tests

The Statistical American Association (ASA) recommends reporting precise *p-values:* https://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108?scroll=top#.XEB0js9Kj0A

However, it is not clear whether it makes sense to report very small *p*-values (e.g., lower than .0001) in disciplines that deal with small amounts of data, and statistical assumptions may not be very accurate.

A good practice may be to report precise values while being careful about their interpretation. An extremely small *p*-value (e.g., .00000001) may not necessarily provide stronger statistical evidence than a larger *p*-value (e.g., .001), in particular when data are messy, samples are not large, and statistical assumptions are not accurate.

Type I error

Incorrectly rejecting the null hypothesis (**false positive**)

Type I error rate = the probability of rejecting the null hypothesis given that it is true.

When $\alpha = .05$, one expects that the Type I error rate of a significance test is 5%.

Type II error

Failing to reject the null hypothesis when it is not true (**false negative**)

Type II error rate = the probability of failing to reject the null hypothesis given that it is not true

The Type II error rate is often denoted as β (beta)

Controlling for the Type II error rate

Significance tests do not provide any guarantee for the Type II error rate

The larger the size of the sample, the lower the Type II error rate is expected to be

It also depends on how large or small the effect of interest (such as the difference between two means) is

Assessing Type II error rates with R

```
N <- 10000 # Number of experiments

m0 <- 100 # mean of null distribution

# population mean and standard deviation
m = 90
sd = 15

n <- 20 # Sample size

alpha <- .05 # Significance (alpha) level
typeII.errors <- 0

for(i in 1:N){
    sample <- rnorm(n, m, sd)
    p <- t.test(sample - m0)$p.value

    if(p > alpha) typeII.errors <- typeII.errors + 1
}

cat("Type II error rate:", typeII.errors / N, "\n")
```

These values will determine the Type II error

Statistical power

It is defined as follows:

power = Pr(reject H_0 | H_1 is true) = 1 - β

Example: If the Type II error rate is $\beta = .3$, then the power is 1 - .3 = .7 or 70%.

We say that a study has high statistical power, if it is expected to reject the null hypothesis (H_0) with a high probability, given that H_1 is true.

Predicting the power of an experiment

1: Assess the minimum expected **effect size** of interest

How small are the differences that we want to detect?

2: Assess the variance in the population of interest.

(This may not be easy)

3: Based on the above parameters, estimate the minimum sample size needed to show the effect of interest.

Small differences and large variance require larger sample sizes

Underpowered studies

Such studies are unlikely to allow the researcher to choose between H_0 and H_1 at the desired significance level (e.g., $\alpha = .05$)

They have high Type II error rates.

Underpowered studies

Main reasons for low statistical power:

Inadequate sample sizes. This is a common problem for many research disciplines (including HCI research).

Failure to identify potential sources of noise or variance and well control the experimental task. This is always a challenge for researchers.

Large measurement error. A good choice of measures and measurements over multiple task replications can reduce this problem.

Reducing sources of variance

Provide clear experimental instructions that are consistent across participants

If **learning effects** are not of interest, provide enough training before the main experiment

Avoid long sessions to minimize **fatigue effects**

Carefully design the experiment through **pilot studies**

Pilot studies (pilot experiments)

Informal small-scale preliminary studies conducted prior to a formal full-scale study:

- Assess the feasibility, time, and cost of a study
- Identify problems (e.g., adverse effects) and refine the experimental procedures
- Predict how large or small the effect of interest is
  Example: How large or small do we expect the difference in IQ scores between two different groups to be?
- Predict the appropriate sample size for the formal study to ensure adequate statistical power.

What is a reasonable power level?

A power of 80% is generally considered as a good level of power. (Type II Error rate = .20)

Compare this with the 95% confidence level that we commonly use (Type I Error rate = .05)

Type I vs. Type II error

Some reasonable questions:

Q1. Why do we allow any errors at all?

Q2. Why are we mostly concerned about Type I errors (false positives) and less about Type II errors (false negatives)?

Note: The discussion in the following slides is based on Toby Mordkoff's course notes:
http://www2.psychology.uiowa.edu/faculty/mordkoff/GradStats/part%201/l.13%2012power.pdf

Q1. Why allowing errors?

It is a necessity!

As the normal distribution goes to infinity in both directions, the bounds of 100% confidence intervals are always $[-\infty, \infty]$

Such confidence intervals are non-informative and useless.

Similarly, if we wanted to eliminate Type I errors ($\alpha = 0$), we would never be able to reject the null hypothesis.

Q2. Why does science focus on Type I errors?

By tradition, **science is conservative**:

By default, it assumes a **no effect** until someone shows an effect (i.e., rejects the null hypothesis).

Failure to reject the null hypothesis does not change our understanding of reality.

In contrast, rejecting a null hypothesis changes this understanding. If such rejection is false (Type I error), then the cost can be high.

Publication bias

Priority has traditionally been given to papers that report on positive results (i.e., ones that confirm an effect)

Why? Because **anybody** can run a lousy, underpowered experiment (e.g., one that does not try to reduce variance) or an experiment on non-effective treatments or techniques that fails to reject the null hypothesis.

Research disciplines try to protected themselves from the influence of lousy experiments. By not publishing such results, they reduce the influence of Type II errors.

Example

Imagine a drug company that advertises the results of some published studies:

"nothing has been shown to work better"
"no side effects above those with placebo"

Both results may be due to Type II errors.

How difficult do you think it is to produce such results?

Note: Based on Toby Mordkoff's course notes:
http://www2.psychology.uiowa.edu/faculty/mordkoff/GradStats/part%201/I.13%2012power.pdf

Dangers of publication bias (1)

But: By not publishing negative results, the influence of Type I errors can become great.

Example: Telepathy

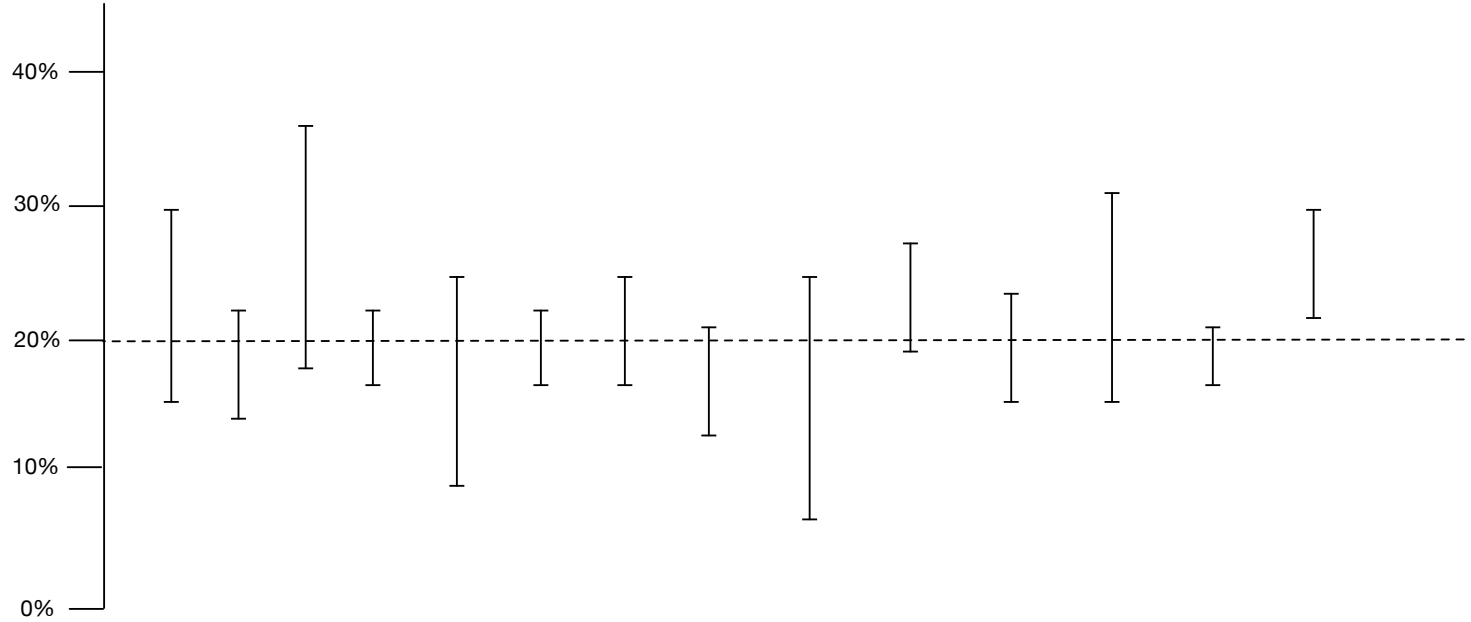
An experiment studies whether two remote people (a sender and a receiver) can communicate information without any physical interaction.

The sender picks a card with a random number from 1 to 5 and tries to communicate it to the receiver. The receiver tries to guess the number communicated by the sender.

> H_0: The receiver has a 20% chance to guess the correct number.
>
> H_1: The receiver guesses the correct number with a rate that is higher than chance (20%).

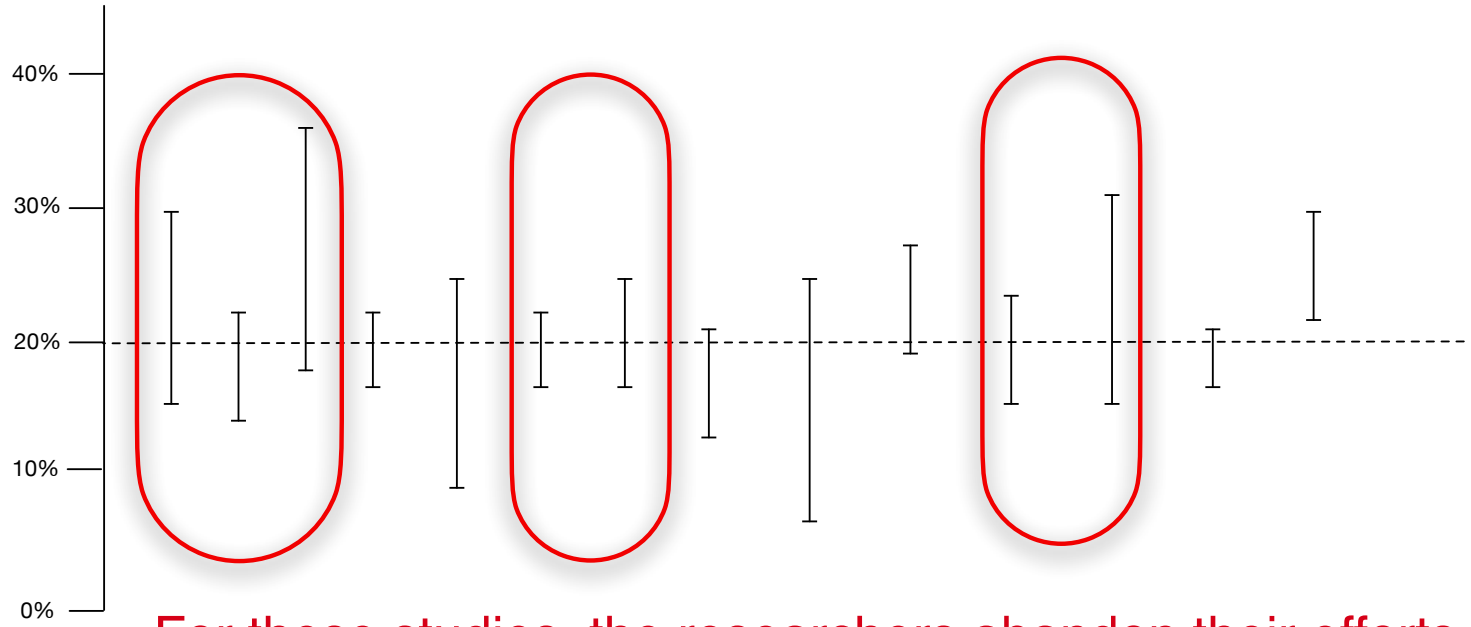
Example: Telepathy

Imagine that different parapsychologists conduct this same experiment many times. The graph below shows the 95% CI of 14 different experiments.



Example: Telepathy

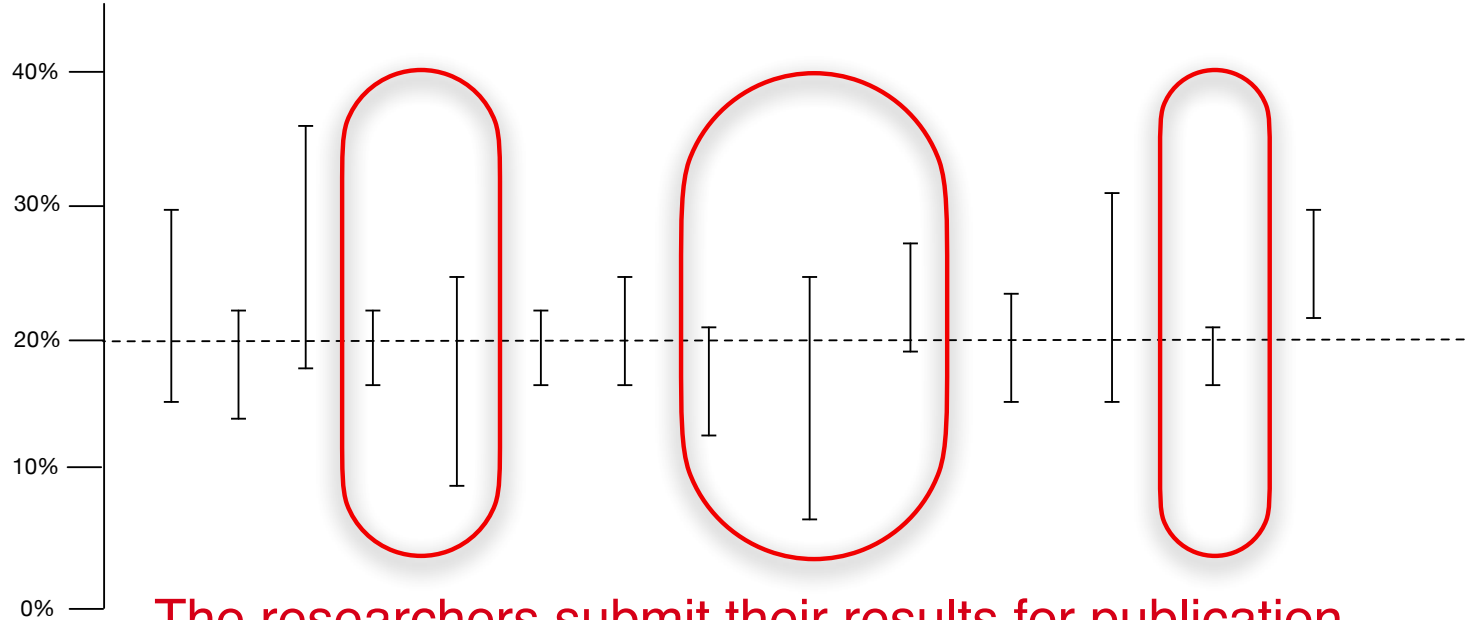
Imagine that different parapsychologists conduct this same experiment many times. The graph below shows the 95% CI of 14 different experiments.



For these studies, the researchers abandon their efforts. They do not even try to publish their results.

Example: Telepathy

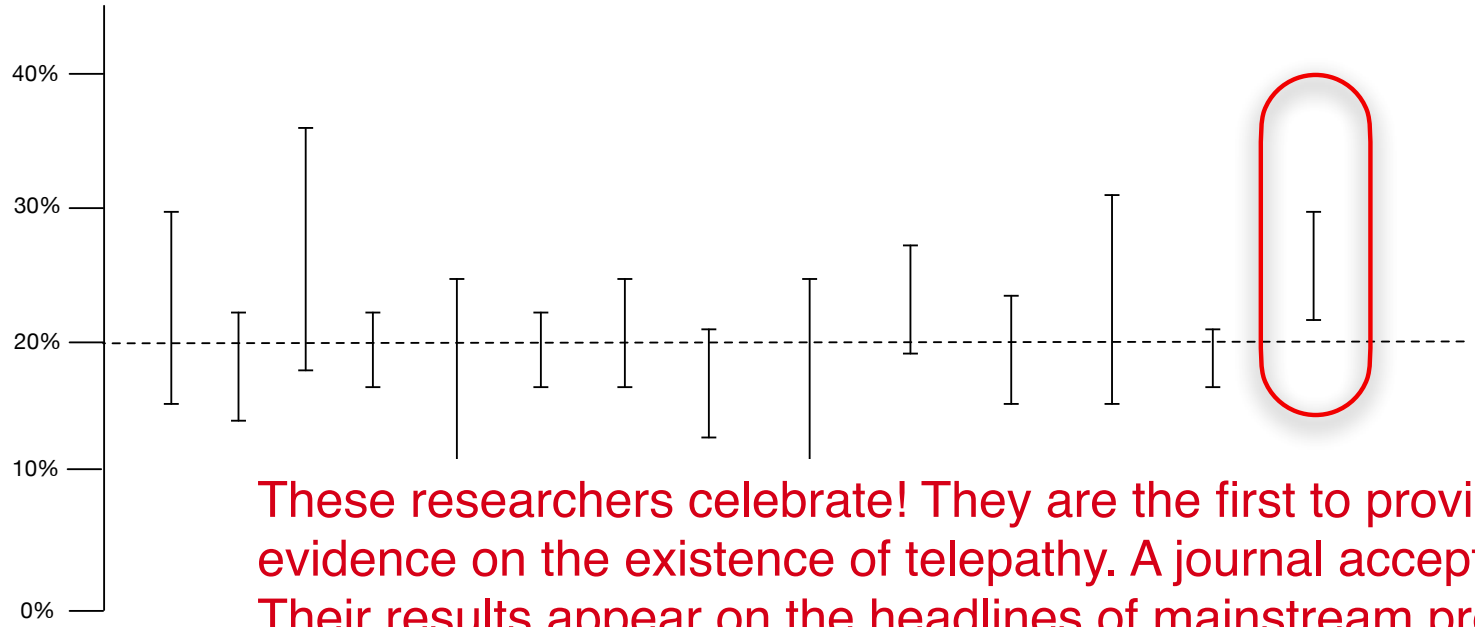
Imagine that different parapsychologists conduct this same experiment many times. The graph below shows the 95% CI of 14 different experiments.



The researchers submit their results for publication.
But all these papers are rejected.

Example: Telepathy

Imagine that different parapsychologists conduct this same experiment many times. The graph below shows the 95% CI of 14 different experiments.



These researchers celebrate! They are the first to provide statistical evidence on the existence of telepathy. A journal accepts their paper. Their results appear on the headlines of mainstream press and trigger many discussions in the social media.

The importance of replication

Such results can be put into question by replicating the experiment. Thus, **publishing the results of replication studies can be important, even if those results are negative.**

In several disciplines, replication is a common practice. Conclusions are often based on **meta-analyses** of the results of several independent studies.

Unfortunately, in other disciplines, replications are rare exceptions.

https://doi.org/10.1136/bmj.323.7327.1450

Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial

BMJ 2001 ; 323 doi: https://doi.org/10.1136/bmj.323.7327.1450 (Published 22 December 2001)

Subjects: All 3393 adult patients whose bloodstream infection was detected at the hospital in 1990-6.

Intervention: In July 2000 patients were randomised to a control group and an intervention group. A remote, retroactive intercessory prayer was said for the well being and full recovery of the intervention group.

Main outcome measures: Mortality in hospital, length of stay in hospital, and duration of fever.

Results: Mortality was 28.1% (475/1691) in the intervention group and 30.2% (514/1702) in the control group (P for difference=0.4). Length of stay in hospital and duration of fever were significantly shorter in the intervention group than in the control group (P=0.01 and P=0.04, respectively).

Conclusions: Remote, retroactive intercessory prayer said for a group is associated with a shorter stay in hospital and shorter duration of fever in patients with a bloodstream infection and should be considered for use in clinical practice.

Remember?

https://doi.org/10.1136/bmj.323.7327.1450

Effects of remote, retroactive intercessory prayer on outcomes in

This study is clearly flawed! The author here has just divided (by chance or intentionally) the control and the intervention group in a convenient way. It is trivial to show the flaw!

Create a script that randomly divides the patients into two groups (control vs. intervention). Repeat it 1000 times. Each time, pray for the intervention group!

For approximately 50 of these repetitions, you will be able to reject the null hypothesis, i.e., you will find that retroactive prayer works!

Note: The author here raises a discussion on scientific methods in a rather controversial manner. He does not suggest that retroactive prayer might work…

and should be considered for use in clinical practice.

nature International weekly journal of science

NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

Don't trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.

In the biggest project of its kind, Brian Nosek, a social psychologist and head of the Center for Open Science in Charlottesville, Virginia, and 269 co-authors repeated work reported in 98 original papers from three psychology journals, to see if they independently came up with the same results.

Brian Nosek's team set out to replicate scores of

Dangers of publication bias (2)

Reaching statistical significance ($p < .05$) has somehow become the ultimate goal of research.

This has also lead to **misuses of significance tests** and unscientific but widely spread practices, commonly known as **p hacking**.

Common *p*-hacking techniques

Incrementally increase the sample size *n* (by recruiting new participants) until statistical significance is achieved.

Selectively remove “outliers” or replace participants that do not conform to the expectations of researchers.

Change the goals or hypotheses based on the results.
A common practice is to focus on random “statistically significant” results, even if these results were not part of the initial experimental goals.

Apply different significance tests and pick the one that leads to significance, often disregarding its underlying assumptions.

p hacking and Type I error

Write an R simulation that estimates the Type I error of a simple *t* test when incrementally increasing the sample size of an experiment.

R Code

```
N <- 10000 # Number of experiments
n <- 12 # Initial sample size
nmax <- 16 # Max sample size. Here, the researcher stops the experiment.

alpha <- .05

error <- 0
for(i in 1:N){
    count <- n
    sample <- rnorm(n)
    p <- t.test(sample)$p.value


     while(p > alpha && count <= nmax){
        sample <- c(sample, rnorm(1))
        p <- t.test(sample)$p.value
        count <- count + 1
     }

    if(p < alpha) error <- error + 1
}

cat("Type I error:", error / N, "\n")
```

Multiverse Analyses: Report on the results of multiple alternative statistical analyses

https://explorablemultiverse.github.io/

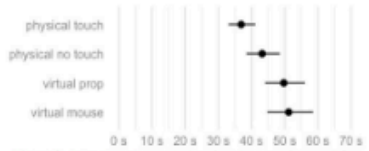
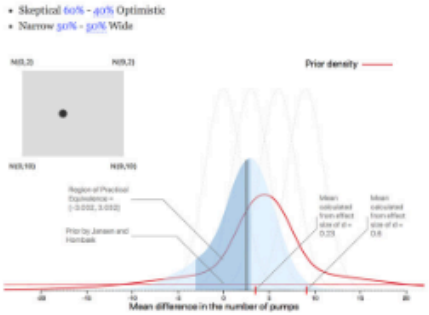


Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CIs.

We focus our analysis on task completion times, reported in Figures 3 and 4. Dots indicate sample means, while error bars are 95% confidence intervals computed on log-transformed data [6] using the t-distribution method. Strictly speaking, all we can assert about each interval is that it comes from a procedure designed to capture the



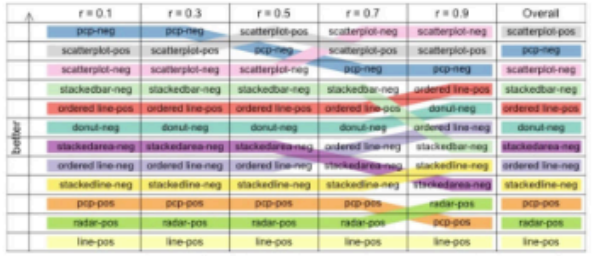


Figure 4. Perceptually-driven ranking of visualizations depending on the correlation sign (-neg / -pos), as a function of correlation value (r) and overall (right column).

Pierre Dragicevic (Inria), Yvonne Jansen (CNRS - Sorbonne Université), Abhraneel Sarma (University of Michigan)
Matthew Kay (University of Michigan), Fanny Chevalier (University of Toronto)

With **explorable multiverse analysis reports**, readers of research papers can explore alternative analysis options by interacting with the paper itself. This new approach to statistical reporting draws from two recent ideas: multiverse analysis, a philosophy of statistical reporting where paper authors report the outcomes of many different statistical analyses in order to show how fragile or robust their findings are; and explorable explanations, narratives that can be read as normal explanations but where the reader can also become active by dynamically changing some elements of the explanation.

Criticisms of the Null Hypothesis Significance Testing (NHST)

Many have recommended abandoning null hypothesis testing altogether and focusing on **estimation** based on **confidence intervals**, **effect sizes**, and **meta-analyses**.

Confidence intervals: Estimate the range of plausible values of a statistic.

Effect sizes: Estimate how large or small an effect is

Meta-analysis: Emphasis on replication and integration of evidence from multiple studies.

Criticisms of NHST

"We need to make two substantial changes to how we carry out research. First, in response to heightened concern that our published research literature is incomplete and untrustworthy, we need new requirements to ensure research integrity. These include **full pre-specification of studies** wherever possible, **avoidance of selection** and other inappropriate data analytic practices, **complete reporting** of research, and encouragement of **replication**. Second, renewed recognition of the many severe flaws of null hypothesis significance testing (NHST) motivates a shift from reliance on NHST to estimation. 'The new statistics' refers to the full range of recommended practices, including **estimation based on effect sizes (ESs), confidence intervals (CIs), and meta-analysis.**"

[Geoff Cumming, 2013]

CIs vs. *p*-values

p-values or 95% CI? What is more informative here?

Given this interval, we can reasonably conclude that the difference between C and A is very likely lower than 1

these p-values are the same and tell us very little!

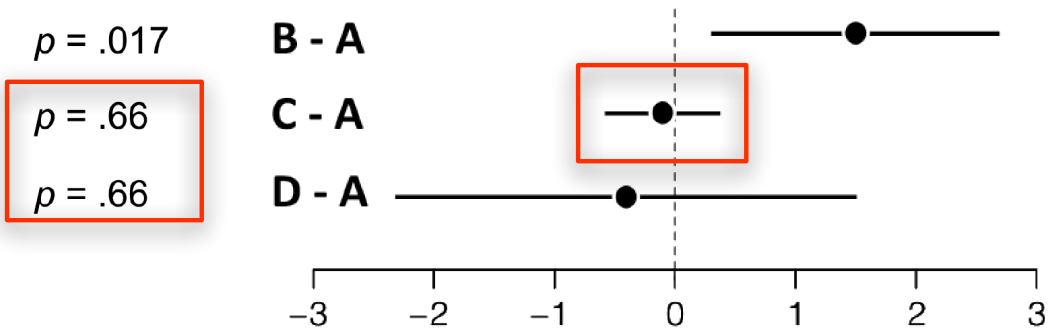


Fig. 6 95% confidence intervals showing differences between conditions.

Fair statistical communication in HCI [Pierre Dragicevic]
https://www.lri.fr/~dragice/fairstats-last.pdf

CIs vs. *p*-values

p-values or 95% CI? What is more informative here?

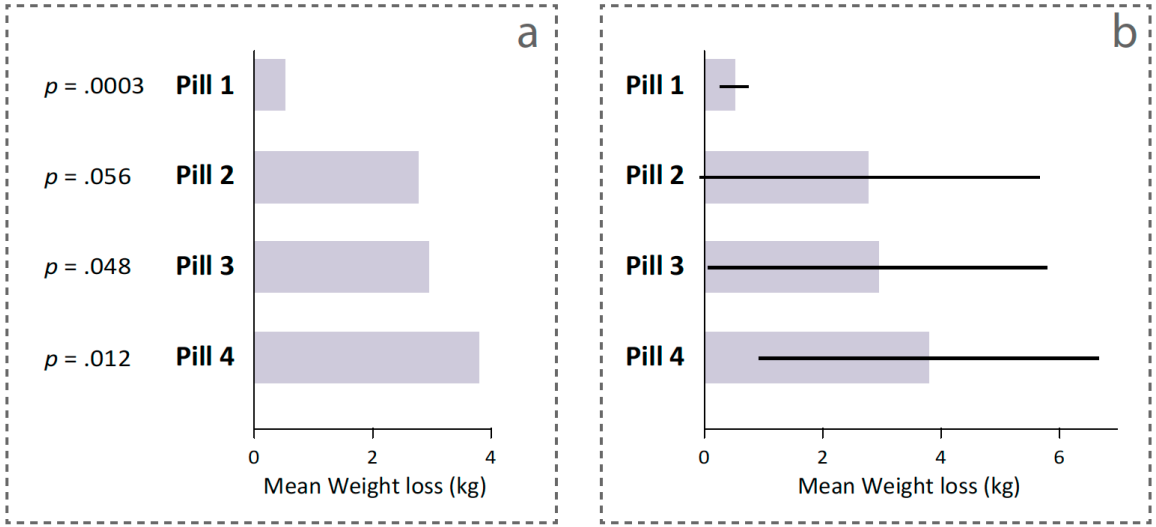


Fig. 4 Showing the most plausible effect sizes and their associated uncertainty using a) *p*-values with point estimates of effect sizes (here shown as bar charts); b) 95% CIs around point estimates.

Fair statistical communication in HCI [Pierre Dragicevic]
https://www.lri.fr/~dragice/fairstats-last.pdf

American Statistical Association (ASA)

ASA's statement on *p*-values (2016):

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Full statement:
http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108?needAccess=true

Interview with ASA's executive director

Ron Wasserstein:

In the post $p<0.05$ era, scientific argumentation is not based on whether a p-value is small enough or not. Attention is paid to effect sizes and confidence intervals. Evidence is thought of as being continuous rather than some sort of dichotomy. (As a start to that thinking, if p-values are reported, we would see their numeric value rather than an inequality ($p=.0168$ rather than $p<0.05$)). All of the assumptions made that contribute information to inference should be examined, including the choices made regarding which data is analyzed and how. In the post $p<0.05$ era, sound statistical analysis will still be important, but no single numerical value, and certainly not the p-value, will substitute for thoughtful statistical and scientific reasoning.

March 2016

Interview text:
http://retractionwatch.com/2016/03/07/were-using-a-common-statistical-test-all-wrong-statisticians-want-to-fix-that/

Why then still use *p*-values?

“ More drastic steps, such as the ban on publishing papers that contain *P*-values instituted by at least one journal, could be counter-productive, says Andrew Vickers, a biostatistician at Memorial Sloan Kettering Cancer Center in New York City.

He compares attempts to ban the use of *P*-values to addressing the risk of automobile accidents by warning people not to drive - a message that many in the target audience would probably ignore. Instead, Vickers says **that researchers should be instructed to *treat statistics as a science, and not a recipe.*** ”

http://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503

Why then still use *p*-values?

”One reason for its continued use is the lack of an agreed alternative. At this stage, it is sufficient to realize that **statistical significance is only one, relatively modest, component in the evaluation of quantitative data.**

This point has been articulated by a number of experts, notably by Abelson (1995). Abelson’s position is that statistics can be viewed as a form of principled argument. Statistical significance, in the form of a conventional NHST, is only one potential element of such an argument.”

[Thom Baguley]

Why then still use *p*-values?

”The main drawback of making the switch to CIs is that it can be difficult to translate some tests into equivalent interval estimates that are both easy to interpret and easy to plot (e.g., tests with multiple degrees of freedom). These situations provide the strongest case for retaining NHSTs.”

[Thom Baguley]

Example

A research team is interested in assessing the effect of a Geometry course on students' IQ performance. They randomly create two groups of students (*Control vs. Geometry*).

Each student takes **three IQ tests over three weeks**.

Control Group

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Week 1	90	88	80	103	91	105	82	91	100	82
Week 2	98	102	93	120	85	105	96	88	105	88
Week 3	107	119	94	112	97	122	104	97	107	92

Geometry Class Group

	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
Week 1	105	94	108	101	99	115	100	95	98	106
Week 2	106	98	110	114	95	122	91	105	92	112
Week 3	107	89	118	106	97	123	102	103	97	111

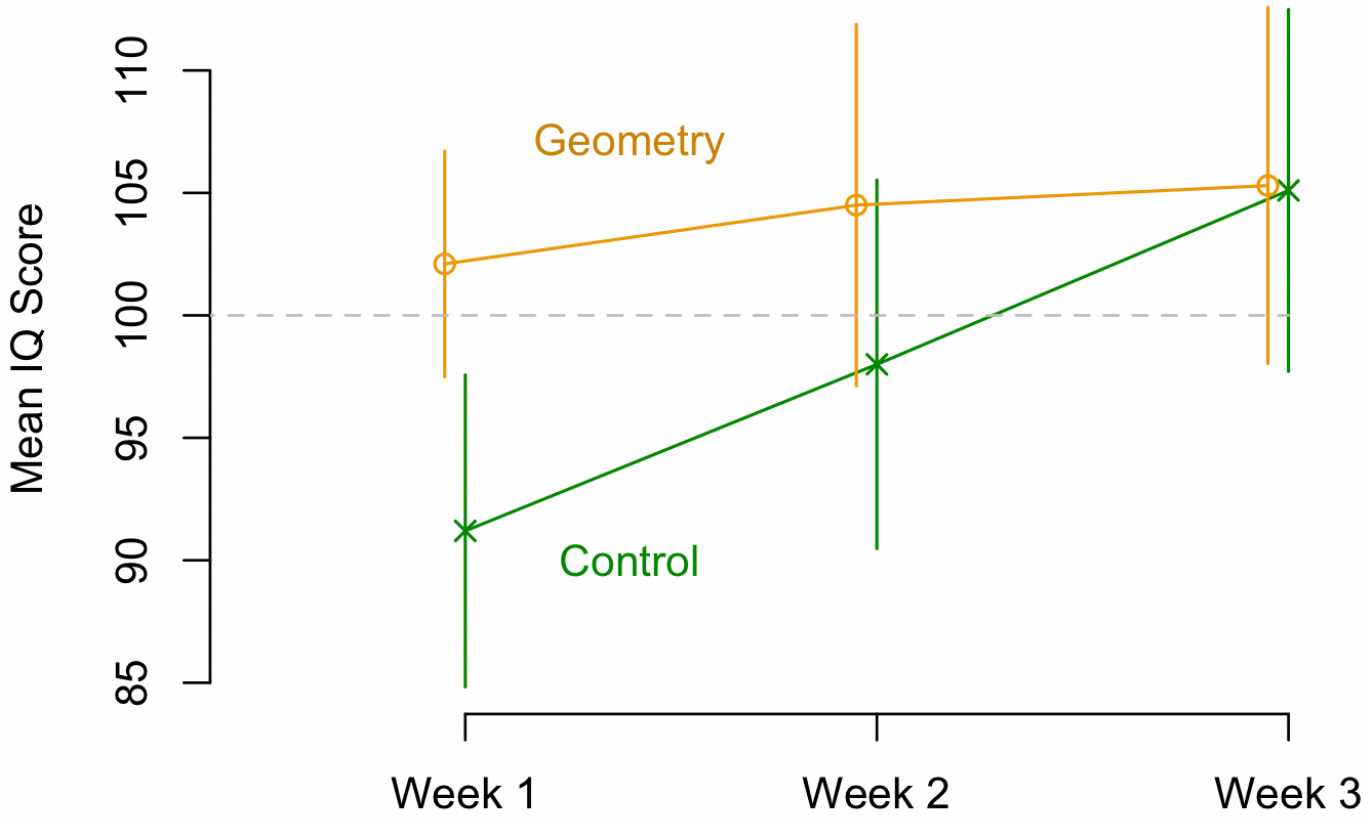
Example

Suppose the researchers are interested in the following questions:

- Is there a difference between the two student groups?
- Does student IQ performance improve with repeated tests?
- Is there are an **interaction** between repeated tests and the student group (Control vs. Geometry)?

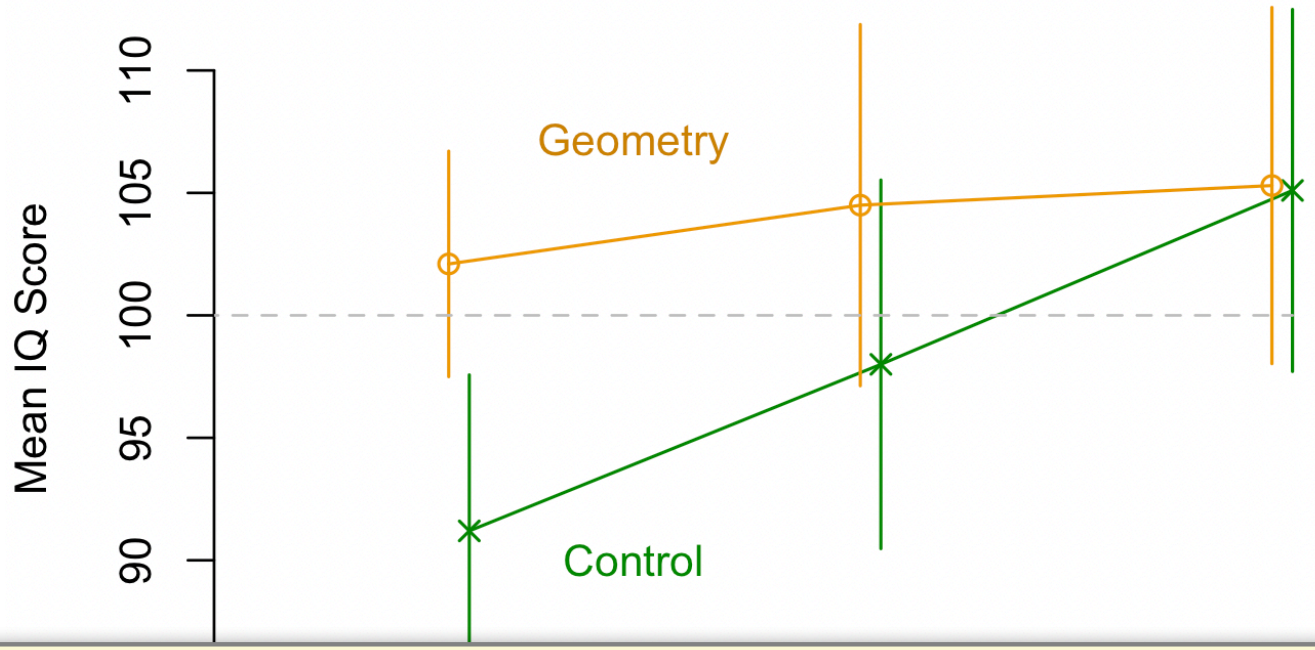
Example

Let's plot the 95% CIs for the mean scores of each group and week.



Example

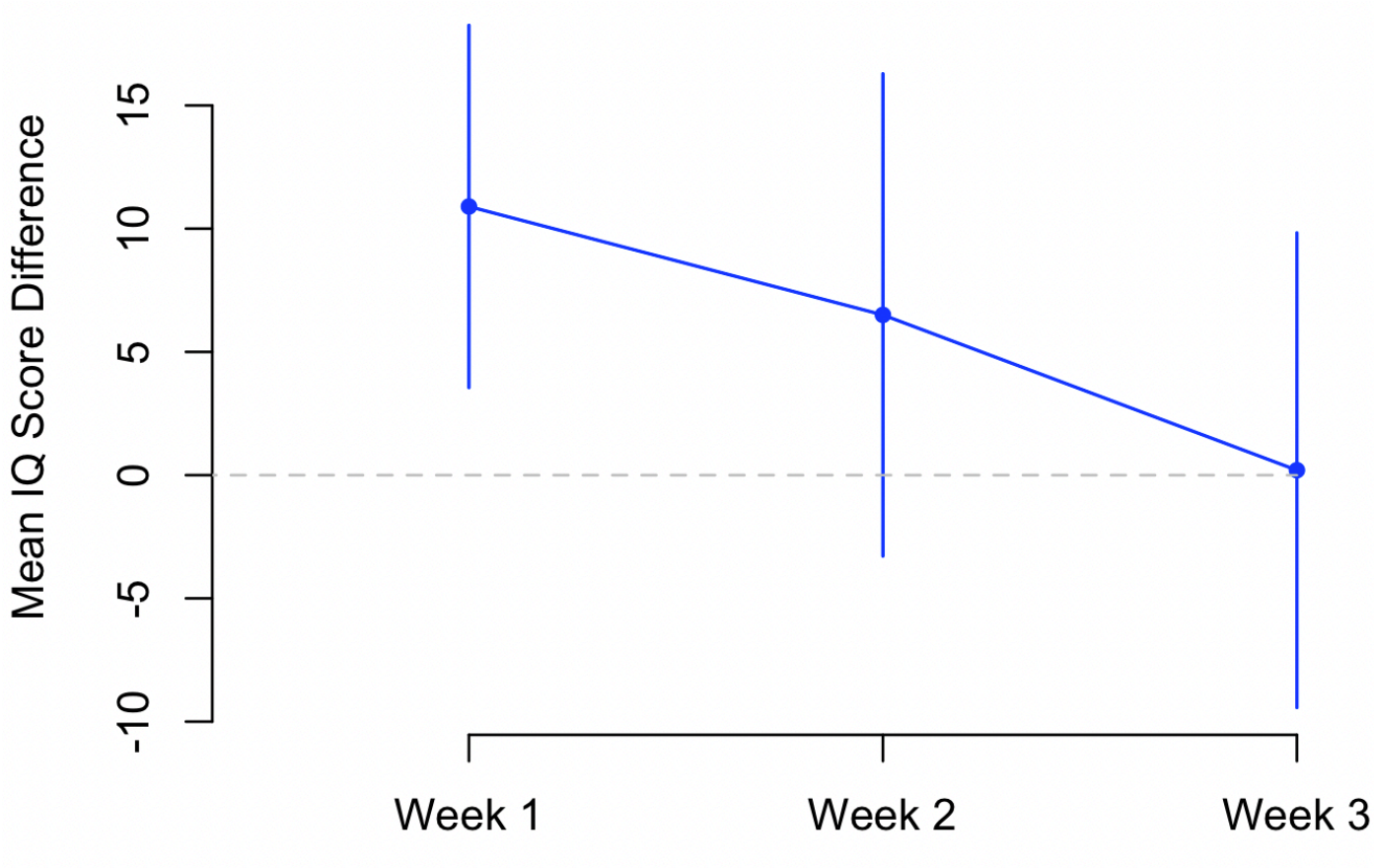
Let's plot the 95% CIs for the mean scores of each group and week.



This is what we call an **interaction effect** between multiple independent variables. There is a simultaneous effect of the two variables here: the effect of repeated tests over several weeks depends on the student group (Geometry or Control).

Example

We can also plot the 95% CIs for the mean score differences.



Example

The analysis becomes more complex if there are more than two independent variables with multiple levels.

In this case, we can combine estimation through confidence intervals with statistical models that can describe complex relationships between variables, such as **Analysis of Variance (ANOVA)** and **Linear** (or Generalized) **Mixed-Effect Models**.

Such models are our of the scope of this course.

Example: an ANOVA model in R

anovaModel <- aov(score ~ group*week + Error(factor(participant)/week), data)

summary(anovaModel)

```
Error: factor(participant)
          Df Sum Sq Mean Sq F value Pr(>F)
group      1    516   516.3   2.306  0.146
Residuals 18   4030   223.9

Error: factor(participant):week
           Df Sum Sq Mean Sq F value   Pr(>F)
week        1  731.0   731.0   32.79 1.98e-05 ***
group:week  1  286.2   286.2   12.84  0.00212 **
Residuals  18  401.2    22.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value for the overall effect of the student group

p-value for the effect of repeated tests

p-value for their interaction

Updated ASA recommendations (2019)

https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913#_i2

2. Don't Say "Statistically Significant"
The ASA Statement on P-Values and Statistical Significance stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

Ronald Fisher (1890 - 1962). He established the notions of the *null hypothesis* and *significance testing*, popularized *p-values,* and developed the *analysis of variance*.

In his earlier works, he used *.05* as a convenient threshold for statistical significance.

> *"The value for which P = 0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not."*
>
> R. Fisher, *Statistical Methods for Research Workers,* 1925

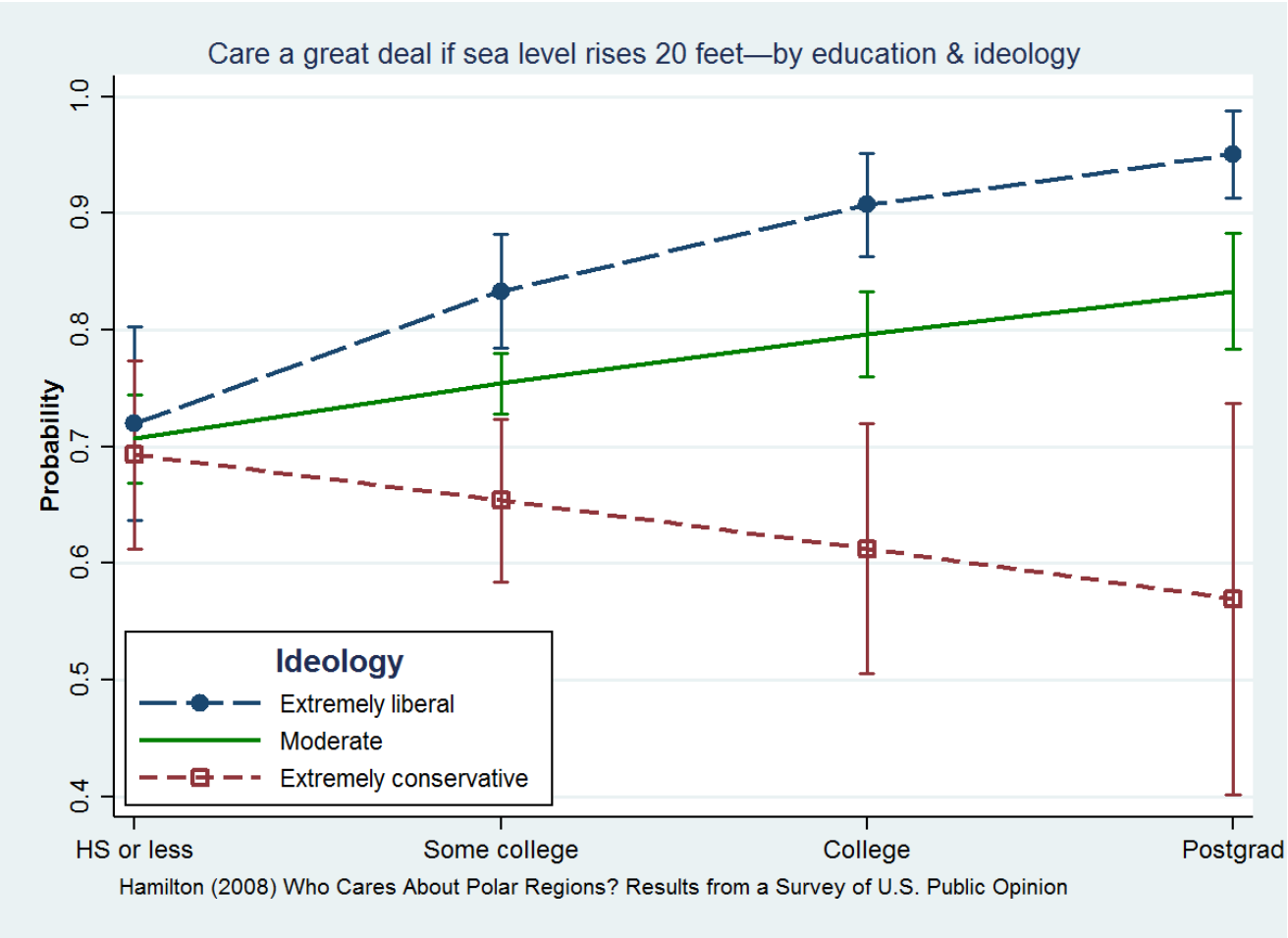
In his later writings, however, he argued that appropriate levels of significance largely depend on the research in question and may vary from research to research. According to his approach, reporting exact *p*-values is important for assessing the level of statistical evidence.

Some argue that modern version of hypothesis testing is an inconsistent hybrid of Fisher's method and the method of Neyman and Pearson (junior), resulting from confusion by writers of statistical textbooks.

A long scientific dispute between Fisher and Neyman-Pearson only ended after Fisher's death.

Interaction effects (further examples)

Interaction effect of education and ideology on concern about sea level rise



from https://en.wikipedia.org/wiki/Interaction_(statistics)

Multiple Comparisons and Registration

Example

Suppose a research team is interested in whether and how listening to music affects the performance of kids in IQ tests. To this end, they test three groups of 10 year old participants:

(G1) No music (20 participants)

(G2) Classical music (20 participants)

(G3) Rock music (20 participants)

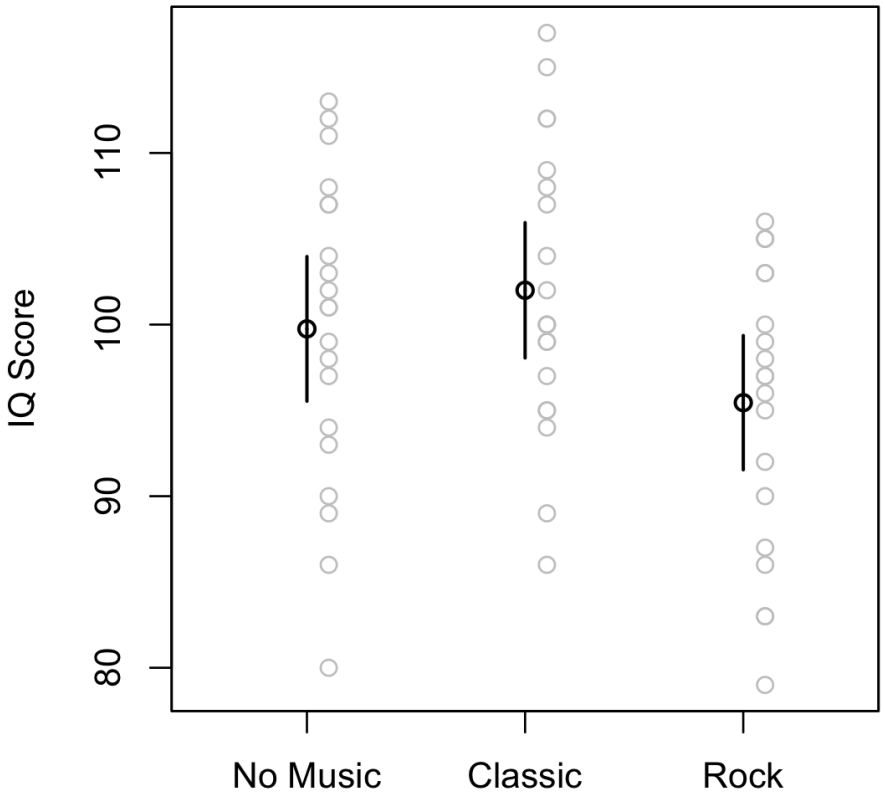
These are the IQ scores for the three groups:

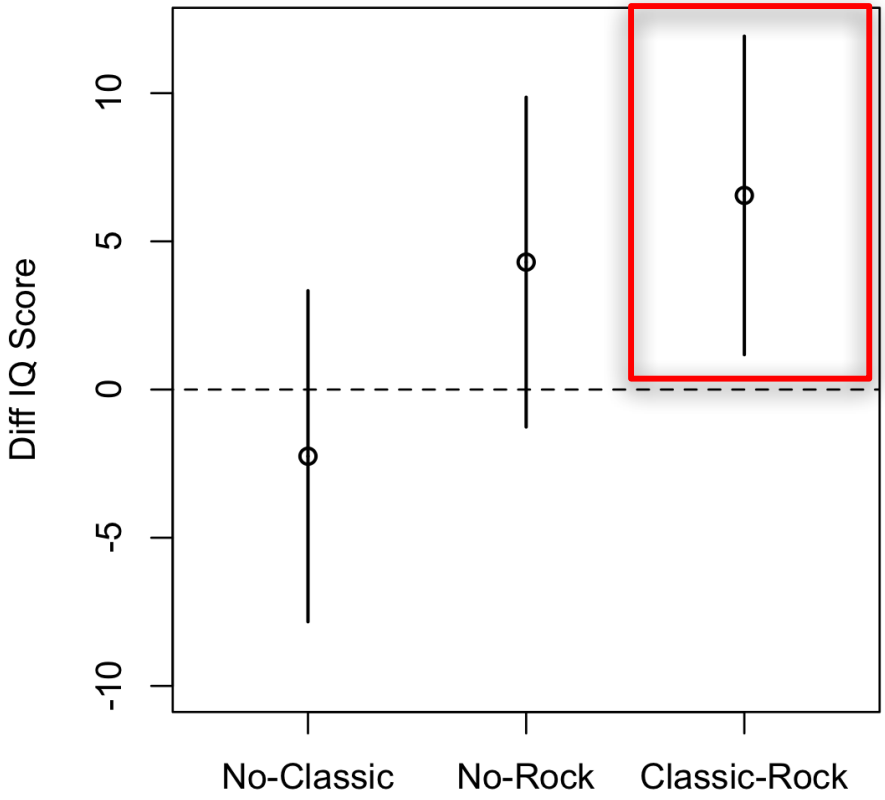
```
G1 <- c(102, 97, 90, 107, 89, 104, 101, 112, 86, 108, 113, 94, 80, 98, 101, 107, 103, 111, 93, 99)

G2 <- c(112, 100, 86, 115, 94, 100, 112, 117, 95, 99, 89, 108, 97, 104, 109, 102, 100, 95, 99, 107)

G3 <- c(105, 83, 92, 106, 98, 86, 103, 87, 97, 103, 83, 90, 105, 79, 100, 96, 97, 105, 95, 99)
```

Summary of results

One could claim here that there is a statistically significant difference (α = .05) between these two groups.

Bars show 95% Confidence Intervals



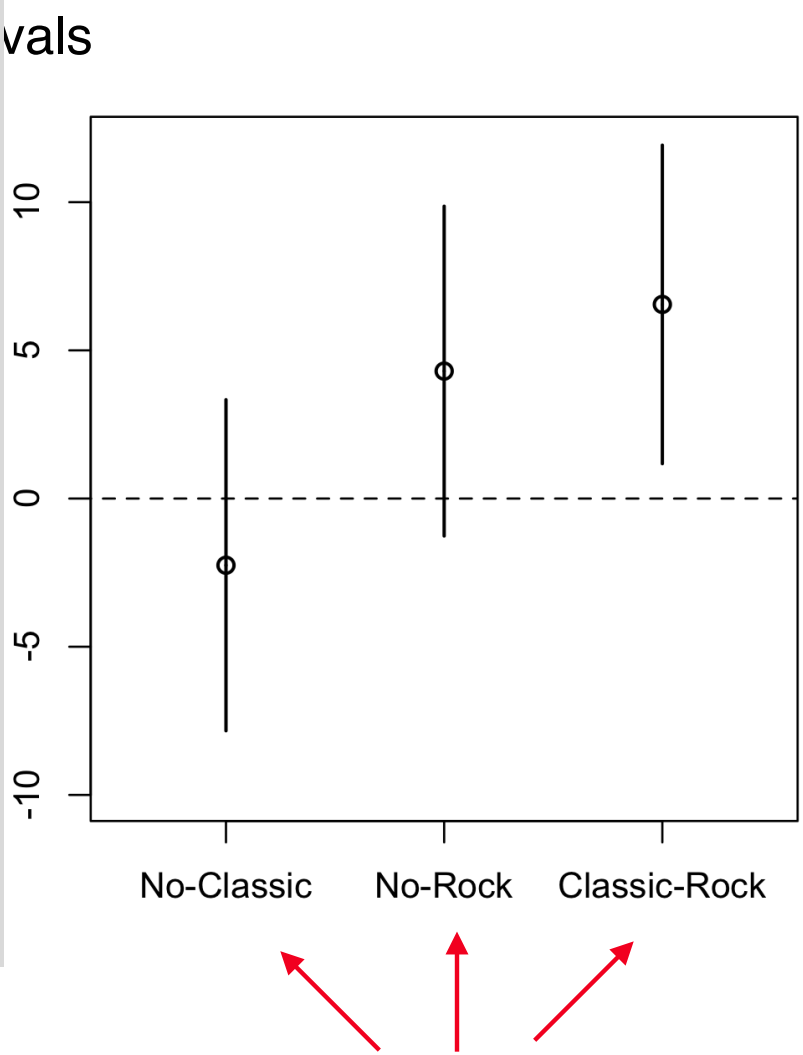


Summary of results

But the researchers perform a total of 3 comparisons here.

Unfortunately, the more the comparisons, the more the danger to find a random difference as statistically significant!

The Type I error rate increases with the number of tests that we do.



Accounting for multiple comparisons

We adjust the α level with respect to the number k of comparisons that we test: $\alpha_k = \alpha / k$

For $k = 3$ comparions and $\alpha = .05$, the corrected level will be $\alpha_3 = 0.05/3 = .0167 = 1.67\%$

Thus, a difference will be assessed as statistically significant if $p < .0167$

(This is commonly known as the **Bonferroni correction**)

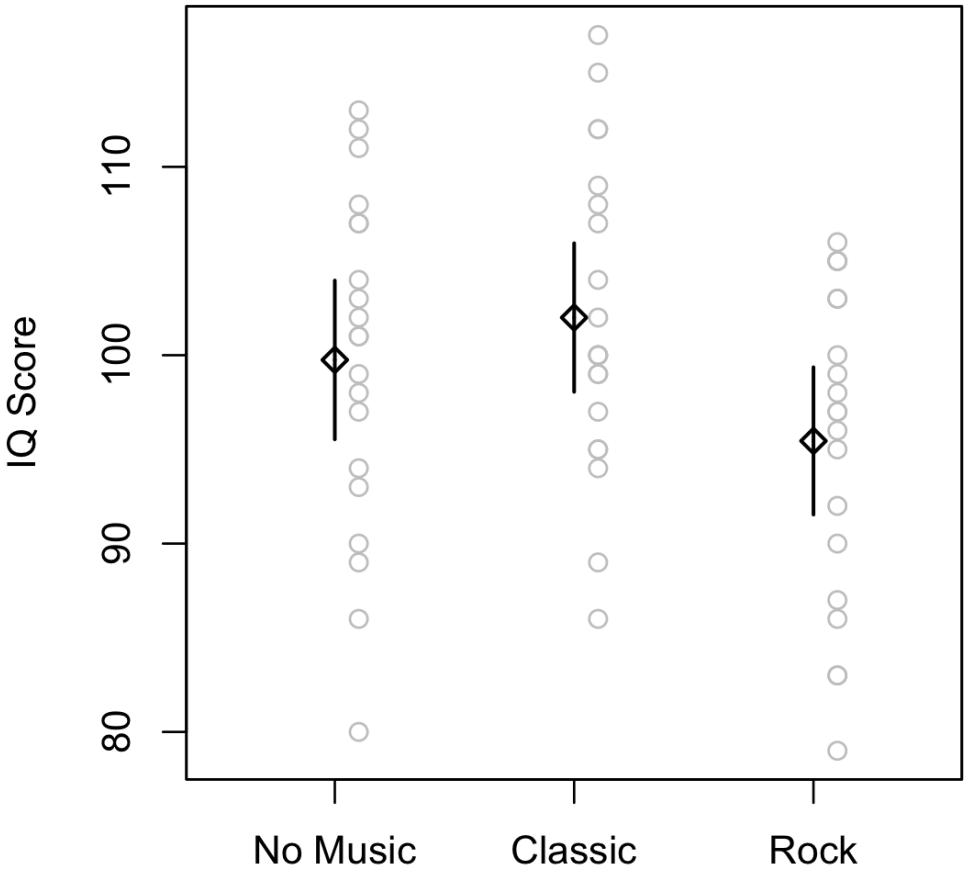
Corrected Confidence Intervals

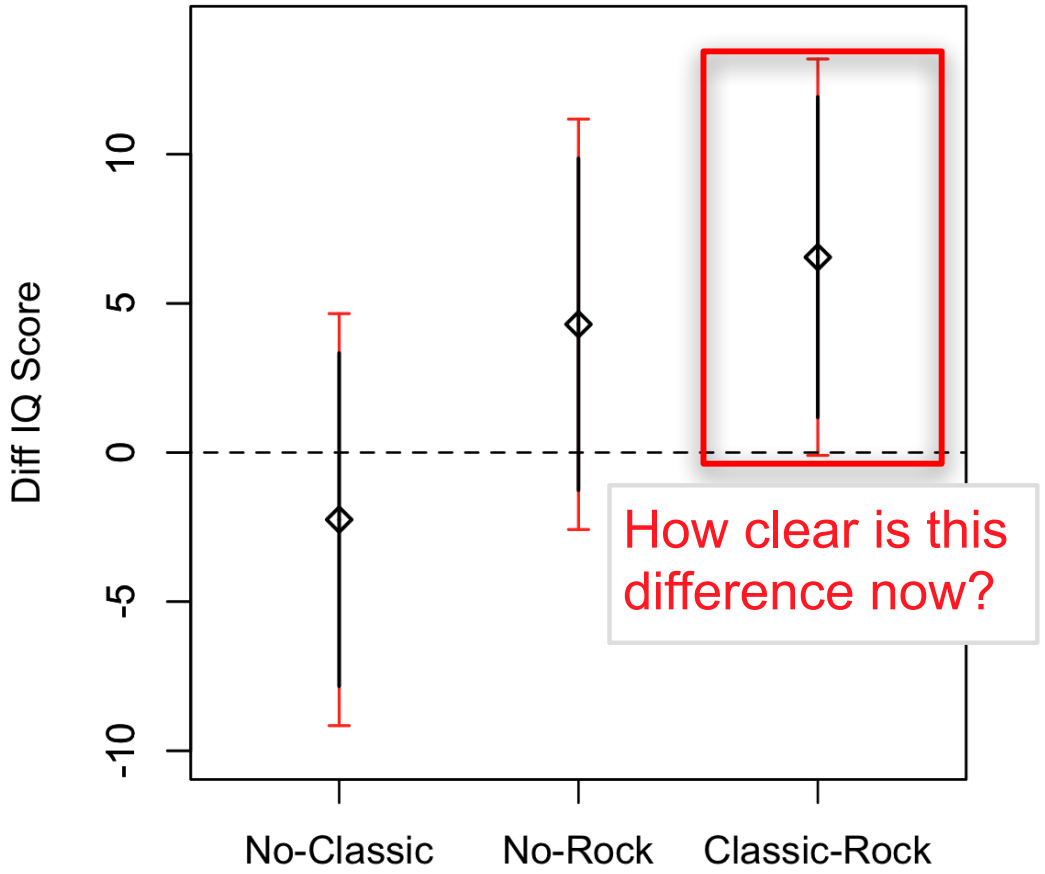
To account for *k* comparions, we will construct *(100 - α/k)% CIs*

For example, for 3 comparisons we will correct the 95% CIs by constructing (100 - 1.67)% = 98.33% CIs

Back to our example

Bars show 95% CIs - **RED** extensions correspond to Bonferroni corrected CIs





Problems

There is no clear consensus on how to correct for multiple comparisons.

Should we also correct for comparisons on different independent variables?

Should we correct for planned tests whose effect has been predicted in advance or only for multiple ***post hoc*** comparisons?

Some good practices

Try to predict the effects based on previous theory, past studies, or early results.

Plan the comparisons of interest in advance.

Assess the power of an experimental design and select sample sizes with respect to the number of planned comparisons.

Keep the number of comparisons low.

In contrast, conducting all possible tests, looking for all possible differences is a bad practice.

Preregistration

Many disciplines and scientific journals have started encouraging (or even requiring) the **preregistration** of research protocols.

Preregistration is a kind of publishing the goals, hypotheses, and methodology (procedures, sample sizes, types of analysis and tests, etc.) of a study before it begins.

The goal is to avoid publication bias and reduce the risk of false positive results, thus reduce the risk of Type I errors.

Preregistration

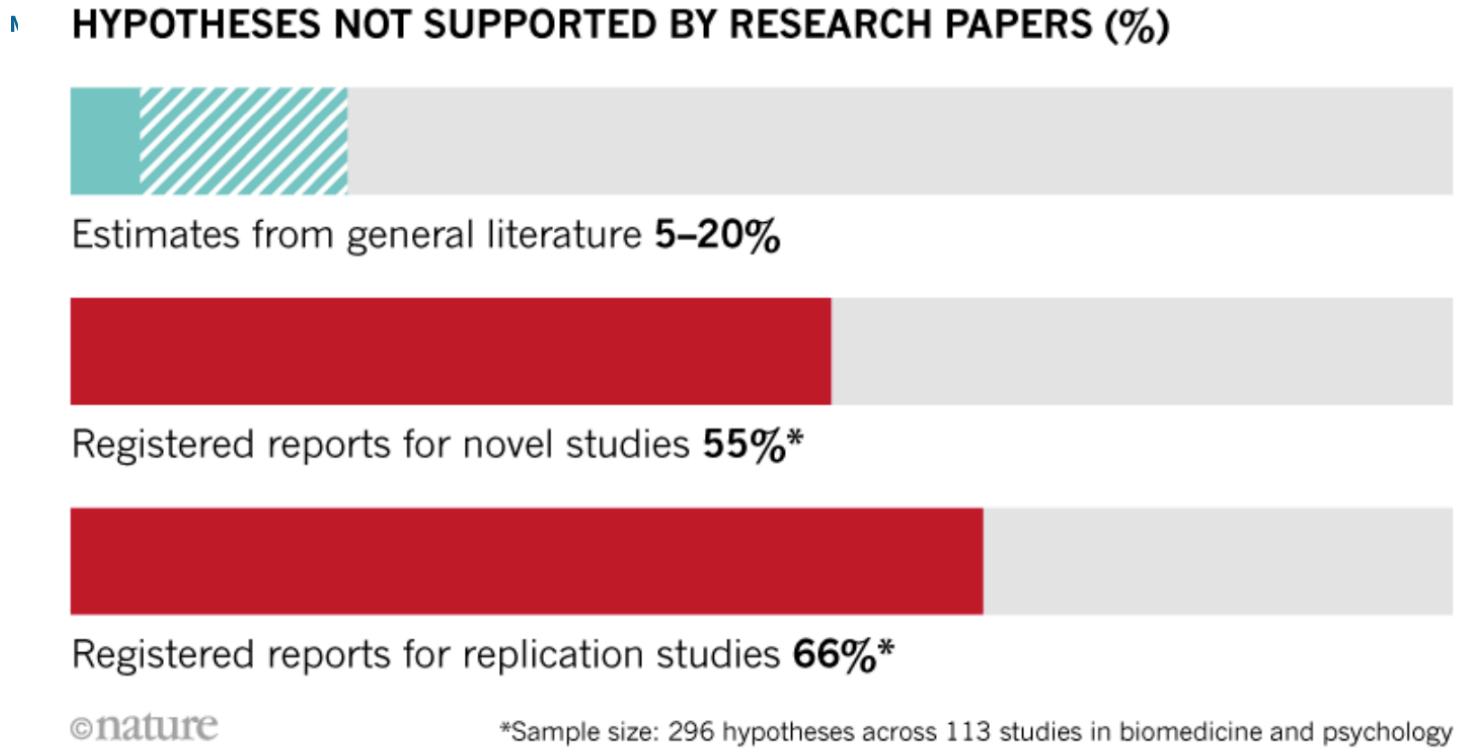
There are several open services for registering research protocols, e.g., see the **Open Science Framework** [https://osf.io/registries]

https://www.nature.com/articles/d41586-018-07118-1

NEWS · 24 OCTOBER 2018

First analysis of 'pre-registered' studies shows sharp rise in null findings

Logging hypotheses and protocols before performing research seems to work as intended: to reduce publication bias for positive results.



Source: Allen, C. & Mehler, D. Preprint at PsyArXiv https://psyarxiv.com/3czyt (2018).

Exploratory Findings vs. Confirmatory Tests

There is a distinction between **confirmatory research** that uses data to test hypotheses and **exploratory research** that uses data to generate hypotheses.

Some studies can be largely exploratory, helping researchers to identify interesting patterns or trends in data that were not evident before running the study.

Researchers should be clear about which parts of their analyses are confirmatory and which parts are exploratory. Results from exploratory analyses can help researchers frame the hypotheses of future studies.