



# Common mistakes in statistics

Or The dangers of the Mouth of Truth



Rink Hoekstra

Open Science and Reproducibility event

28<sup>th</sup> of March 2018



# Introduction





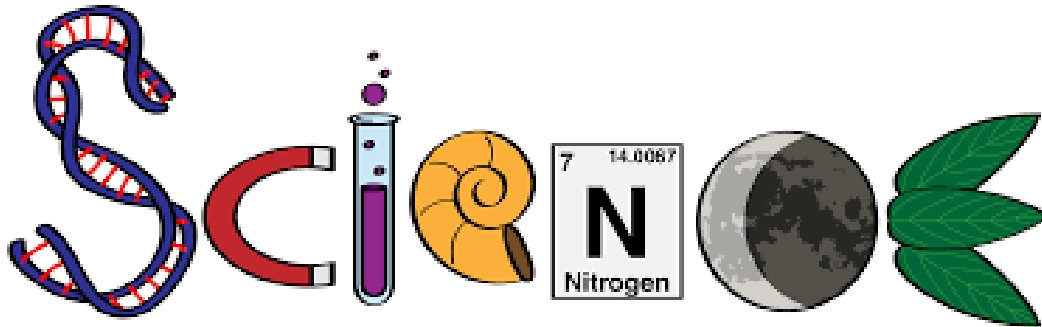
## Structure of this workshop

- Personal introduction
- › Why do we need science?
- › Why do we need statistics?
- › Common misinterpretations
- › Now what?



# Why do we need science?

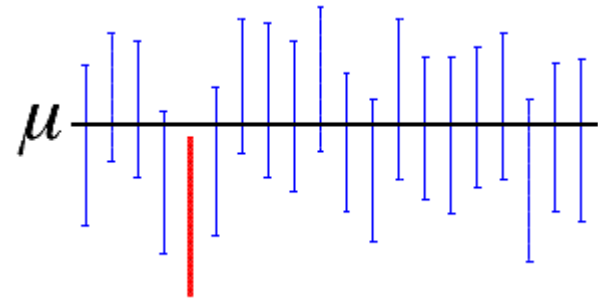
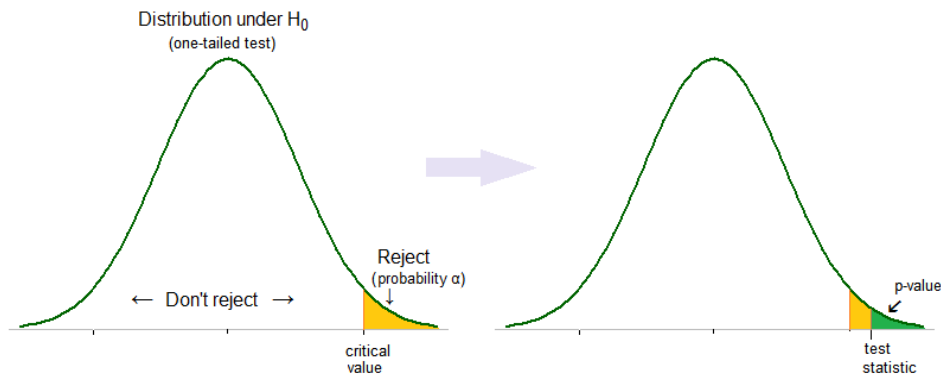
- › What do we expect from science?
- › Why do we need science in the first place?
- › What are we after?





## How does this relate to statistics?

- › How do, for example,  $p$ -values and CIs, help us “finding truth” or “increase knowledge”?







## Let's test our intuitions





## Test about p-values

- › Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means  $t$ -test and your result is significant ( $t = 2.7$ ,  $d.f. = 18$ ,  $p = 0.01$ ). Please mark each of the statements below as “true” or “false.” “False” means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

- › Taken from Gigerenzer 2004



Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means  $t$ -test and your result is significant ( $t = 2.7$ ,  $d.f. = 18$ ,  $p = 0.01$ ). Please mark each of the statements below as “true” or “false.” “False” means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

1. You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).







Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means  $t$ -test and your result is significant ( $t = 2.7$ ,  $d.f. = 18$ ,  $p = 0.01$ ). Please mark each of the statements below as “true” or “false.” “False” means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

2. You have found the probability of the null hypothesis being true.





Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means  $t$ -test and your result is significant ( $t = 2.7$ ,  $d.f. = 18$ ,  $p = 0.01$ ). Please mark each of the statements below as “true” or “false.” “False” means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).





Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means  $t$ -test and your result is significant ( $t = 2.7$ ,  $d.f. = 18$ ,  $p = 0.01$ ). Please mark each of the statements below as “true” or “false.” “False” means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

4. You can deduce the probability of the experimental hypothesis being true.





Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means  $t$ -test and your result is significant ( $t = 2.7$ ,  $d.f. = 18$ ,  $p = 0.01$ ). Please mark each of the statements below as “true” or “false.” “False” means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

5. You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.







Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means  $t$ -test and your result is significant ( $t = 2.7$ ,  $d.f. = 18$ ,  $p = 0.01$ ). Please mark each of the statements below as “true” or “false.” “False” means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

TRUE  
 FALSE





## Another intuition

- › How reliable is a single  $p$ -value? What would happen if we did the same study again?

- › [The dance of the  \$p\$ -values](#)





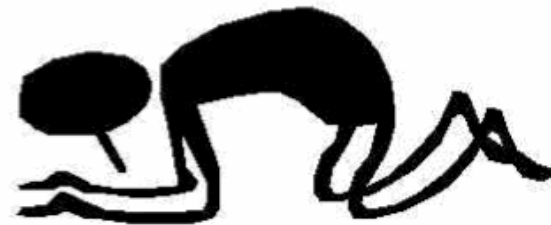
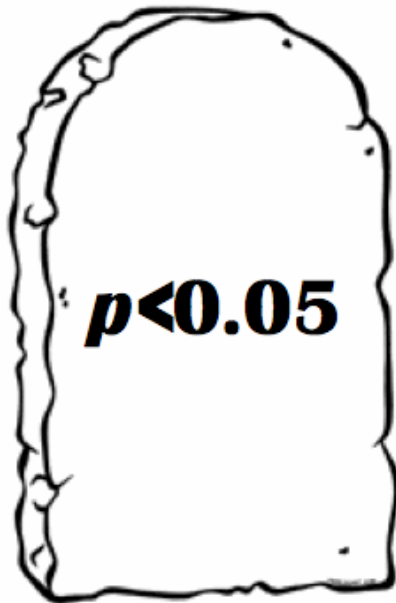
## Yet another intuition

- › Significance testing as a magical truth telling machine





# What does “significant” actually mean?





## Statements about p-values

1. You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).
2. You have found the probability of the null hypothesis being true.
3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
4. You can deduce the probability of the experimental hypothesis being true.
5. You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision
6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

*NB: neither of the statements do logically follow*





# Confidence intervals

Professor Bumbledorf conducts an experiment, analyzes the data, and reports:

The 95% confidence interval  
for the mean ranges from 0.1  
to 0.4!



---

Please mark each of the statements below as “true” or “false”. False means that the statement does not follow logically from Bumbledorf’s result. Also note that all, several, or none of the statements may be correct:

Taken from Hoekstra, Morey, Rouder & Wagenmakers, 2014





Professor Bumbledorf conducts an experiment, analyzes the data, and reports:

The 95% confidence interval  
for the mean ranges from 0.1  
to 0.4!



Please mark each of the statements below as “true” or “false”. False means that the statement does not follow logically from Bumbledorf’s result. Also note that all, several, or none of the statements may be correct:

I. The probability that the true mean is greater than 0 is at least 95 %





Professor Bumbledorf conducts an experiment, analyzes the data, and reports:

The 95% confidence interval  
for the mean ranges from 0.1  
to 0.4!



Please mark each of the statements below as “true” or “false”. False means that the statement does not follow logically from Bumbledorf’s result. Also note that all, several, or none of the statements may be correct:

2. The probability that the true mean equals 0 is smaller than 5 %.





Professor Bumbledorf conducts an experiment, analyzes the data, and reports:

The 95% confidence interval  
for the mean ranges from 0.1  
to 0.4!



Please mark each of the statements below as “true” or “false”. False means that the statement does not follow logically from Bumbledorf’s result. Also note that all, several, or none of the statements may be correct:

3. The “null hypothesis” that the true mean equals 0 is likely to be incorrect.





Professor Bumbledorf conducts an experiment, analyzes the data, and reports:

The 95% confidence interval  
for the mean ranges from 0.1  
to 0.4!



Please mark each of the statements below as “true” or “false”. False means that the statement does not follow logically from Bumbledorf’s result. Also note that all, several, or none of the statements may be correct:

4. There is a 95 % probability that the true mean lies between 0.1 and 0.4.







Professor Bumbledorf conducts an experiment, analyzes the data, and reports:

The 95% confidence interval  
for the mean ranges from 0.1  
to 0.4!



Please mark each of the statements below as “true” or “false”. False means that the statement does not follow logically from Bumbledorf’s result. Also note that all, several, or none of the statements may be correct:

5. We can be 95 % confident that the true mean lies between 0.1 and 0.4.







Professor Bumbledorf conducts an experiment, analyzes the data, and reports:

The 95% confidence interval  
for the mean ranges from 0.1  
to 0.4!



Please mark each of the statements below as “true” or “false”. False means that the statement does not follow logically from Bumbledorf’s result. Also note that all, several, or none of the statements may be correct:

6. If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4.





## Statements for confidence intervals

1. The probability that the true mean is greater than 0 is at least 95 %.
2. The probability that the true mean equals 0 is smaller than 5 %.
3. The “null hypothesis” that the true mean equals 0 is likely to be incorrect.
4. There is a 95%probability that the true mean lies between 0.1 and 0.4.
5. We can be 95 % confident that the true mean lies between 0.1 and 0.4.
6. If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4.

*NB: neither of the statements do logically follow*



## So how *do* we interpret these outcomes?

### › Meaning $p$ -value:

- “The probability to find this finding or more extreme, provided that  $H_0$  is true”
- Formally:  $P(\geq \text{data} \mid H_0)$
- NB:  $\neq P(H_0 \mid \text{data})$  !!

### › Meaning statistical significance:

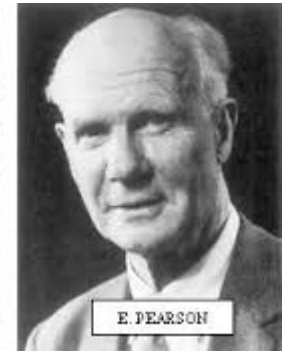
- “We found a  $p$ -value that was smaller than a predetermined significance level. Therefore either we found an unlikely event, or the null hypothesis is not true to begin with”



## Historically

› Debate in the 30s with two *competing* theories about significance testing.

› The Neyman Pearson approach



› The Fisher approach







› Neyman/Pearson approach:

- Compare p-value with alpha
- Make a *decision* about rejecting or accepting (!) the null hypothesis
- No *conclusion* drawn
- Alpha determines error rate in the long run



› Fisher approach

- Draw *conclusions* from the p-value
- Researchers believed to be knowledgeable enough to make sound judgements
- No clear mechanism between outcomes and conclusion







- › The current approach (a.k.a. Null Hypothesis Significance Testing) is a strange hybrid form
  - *Decision* replaced by *conclusion*
  
- › That is, no philosophical foundation for this use
  
- › Philosophically sound interpretation:
  - Neyman/Pearson: Only present *decision*
  - Fisher: present *conclusion* (without *decision*)



## Discrepancy

- › Misinterpretations of  $p$ -values to be found in many papers (even ignoring QRPs, publication bias)
- › This affects its usefulness as a tool to find “truth” or increase knowledge
- › Clearly, the ideals for science and the use of statistics are at odds





## General tendency misinterpretations

- › Simplification outcomes
- › Ignoring uncertainty
- › Convinced, rather than nuanced





## What causes this discrepancy?

- › Is it ignorance of statistical techniques?
- › Or is it adjusting to a perverse incentive structure?  
(stronger stories are more often published)





## Now what?

- › Some suggestions regarding significance testing
  - “Significant” doesn’t mean too much
  - Non-significant does not entail that  $H_0$  is (probably) true
  - Be weary of strong claims, with little or no uncertainty
  - Remember the dance, and The Mouth of Truth



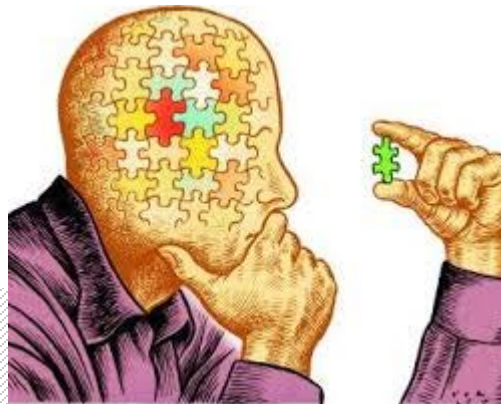


## General suggestions

- Think for yourself
- Be a critical, [independent](#) researcher
- Read!



- › For every decision, try to understand *why* you do it
  - Why (don't) talk about effect size?
  - Why (not) focus on significance testing?
  - Why (not) use Bayesian statistics?
  - Why add/leave out confidence intervals?
  - Does the technique answer my research question?





- › If the answer is
  - “because everybody else is doing it”
  - “because X told me to do it”
  - “statisticians say so, although I haven’t got a clue why”
  - “I only know it increases my chances of getting the paper published”
  
- › ... realize that this may well be at odds with what science should be about





› Of course

- You cannot be *purely* idealistic
- You cannot know everything about every issue

› A good scientist is at least aware *when* he/she is compromising, and what he/she doesn't know





## In summary

- › Think (more) about what you do, and why
- › Read about issues/talk to others
- › Don't accept advice before understanding it  
(including everything I've said)



The supreme lesson  
of education is to  
think for yourself;  
absent this attainment,  
education creates  
dangerous, stupefying  
conformity.

— BRYANT MCGILL

Simple Reminders  
SIMPLEREMINDERS.COM





# Reading suggestions

## Misunderstanding statistics

Cohen, J. (1994). The earth is round ( $p < .05$ ). *American psychologist*, 49(12), 997.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5), 587-606.

## Scientific practice

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5), 524-532.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11), 1359-1366.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

## Bayes

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2016). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic bulletin & review*, 1-16.

## Confidence intervals

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review*, 21(5), 1157-1164.





# Questions?

