Basic statistics for scientists

Jana Hozzová

Institute of Computer Science

Usual routine

- results should look like this
- guess the sample size
- run some experiments
- calculate statistics, until it looks good
- conclude that your idea is correct
- if no "good" results, run more experiments
- maybe this parameter also matters
- maybe you found a bug

More or less okay for exploratory research, but ...

What should not happen

- results should look like this
- guess the sample size
- run some experiments
- calculate statistics, until it looks good
- conclude that your idea is correct
- if no "good" results, run more experiments
- maybe this parameter also matters
- maybe you found a bug

We will talk about ...

- results should look like this -> **Hypotheses**
- guess -> **Sample Size Justification**
- run experiments -> **Pre-registration and Error Control**
- calculate, until ... -> **Calculating statistics and Multiple Testing**
- conclude -> **Interpreting statistics**
- ..., so run more experiments -> **Optional stopping**
- maybe this also matters -> **Dependent and Independent Variables, Hypotheses**

Hypotheses

- the null hypothesis H0: there is nothing is going on
- the alternative hypothesis H1: there is something is going on
- fully formed before data collection
- what experiment could confirm it?
- what experiment could disprove it?

What can happen?

	H0 True	H1 True
Significant Finding	False Positive (α)	True Positive (1-β)
Non-Significant Finding	True Negative (1- α)	False Negative (β)

Statistically significant?

- is this random noise or something?
- calculate test statistic
- calculate the critical value (based on α, distribution, parameters of the test, etc.)
- if the result $<$ critical value $\rightarrow$ statistically significant result
- if the result $>$ critical value $\rightarrow$ not statistically significant result

Test statistic

Take data, calculate one scalar that describes how they differ.

- based on mean: z-test, t-test, f-test
- based on variance: correlation, ANOVA, χ squared test
- based on predicting the change: regression
- non-parametric
- and many more (Wikipedia has 101 pages under category Statistical tests)

What can happen?

	H0 True	H1 True
Significant Finding	False Positive (α)	True Positive (1-β)
Non-Significant Finding	True Negative (1- α)	False Negative (β)

True negative

- when H0 is true, we are saying H0 is true
- important, but almost impossible to publish

True positive

- when H1 is true, we are saying H1 is true
- what we need
  - come up with H1 that are true in this universe
  - have enough statistical power to be able to see it
- statistically significant or not is binary
- the size of the effect

Effect size

- assuming H1 is true, how big is the effect?
- the practical significance of the results
- Cohen's d: what is the difference between x and y?
- correlations: how much the relationship between x and y reduces the error in data?
- you need high statistical power to observe small effect

False negative

- when H1 is true, we are saying H0 is true
- Type 2 error rate β
- in the long run, this error occurs $\beta\%$ of the time
- usually $\beta = 0.2$
- you can decrease it by
  - increasing sample size
  - decreasing measurement error
  - predicting the direction of the effect
- potentially even more dangerous than Type 1 error

False positive

- when H0 is true, we are saying H1 is true
- Type 1 error rate α
- in the long run, this error occurs $\alpha\%$ of the time
- substantially decreased if we replicate studies
- usually $\alpha = 0.05$
- inflates in case of
  - multiple tests with the same data
  - optional stopping

Multiple testing

- data
- three statistical tests
- each of them $\alpha = 0.05$
- what is the total Type 1 error rate?

Multiple testing

- data
- three statistical tests
- each of them $\alpha = 0.05$
- what is the total Type 1 error rate?
- $1 - 0.95^3 = 0.14$
- with 50 tests, it's above 90%
- control with $\alpha = 0.05/3$

Optional stopping

- willing to collect 100 data points
- will look at data three times
- each test has $\alpha = 0.05$
- what is the total Type 1 error rate?

Optional stopping

- willing to collect 100 data points
- will look at data three times
- each test has $\alpha = 0.05$
- what is the total Type 1 error rate?
- ~ 0.1
- with 100 looks, it's ~ 0.35
- control with $\alpha = 0.05/3$

Tricky question

It is known that an effect exists in the population.
In a pilot study, a difference between groups was observed.
Group 1: $n = 22$, $M = 5.68$, $SD = 0.98$.
Group 2: $n = 23$, $M = 6.28$, $SD = 1.11$.
$p < 0.05$

When you replicate the study in a same way **what is the chance you will observe a statistically significant result**?

What can happen?

	H0 True	H1 True
Significant Finding	False Positive (α)	True Positive (1-β)
Non-Significant Finding	True Negative (1- α)	False Negative (β)

We will talk about ...

- results should look like this -> Hypotheses
- guess -> **Sample Size Justification**
- run experiments -> **Pre-registration** and Error Control
- calculate, until ... -> Calculating statistics and Multiple Testing
- conclude -> **Interpreting statistics**
- ..., so you run more experiments -> Optional stopping
- maybe this also matters -> **Dependent and Independent Variables, Hypotheses**

Sample size justification

- according to required accuracy
  - distribution and standard error
- according to required statistical power
  - $1 - \beta$, α and estimated effect size
  - example in RStudio
- according to available resources

Pre-registration

Plan the analysis

- justify sample size
- what data do you measure and how?
- how do you randomize between groups?
- what statistical tests you want to do?
- what are dependent and independent variables for each test?
- specify parameters for each test
- specify error rates α and β and their control

Run and analyze

- run experiments
- collect measurements
- run analysis as planned
- you CAN NOT use data to both come up with and confirm a hypothesis
- exploratory vs. confirmatory research

Interpreting p-values

- we found $p < 0.05$, so ...
- they separate signal from the noise in the data
- is it random variation or is there true difference?
- how much surprising the data are, assuming H0 is true
- example in RStudio

p-value distribution

example in RStudio

p-value < 0.05

it DOES NOT mean

- that H1 is true
- that H1 is probably true (with 1-α probability)
- you will get the same result in replication study
- that your Type 1 error rate is really 0.05
- the value of p depends on the size of the effect

p-value < 0.05

it DOES mean

- you have set your Type 1 error rate to 5%
- that if H0 was true it would be unlikely to observe these data
- if H0 was true we get this result in 5% studies
- somebody should do a replication study to rule out Type 1 error
- that if H1 is true, lower values of p are more probable

p-value > 0.05

it DOES NOT not mean

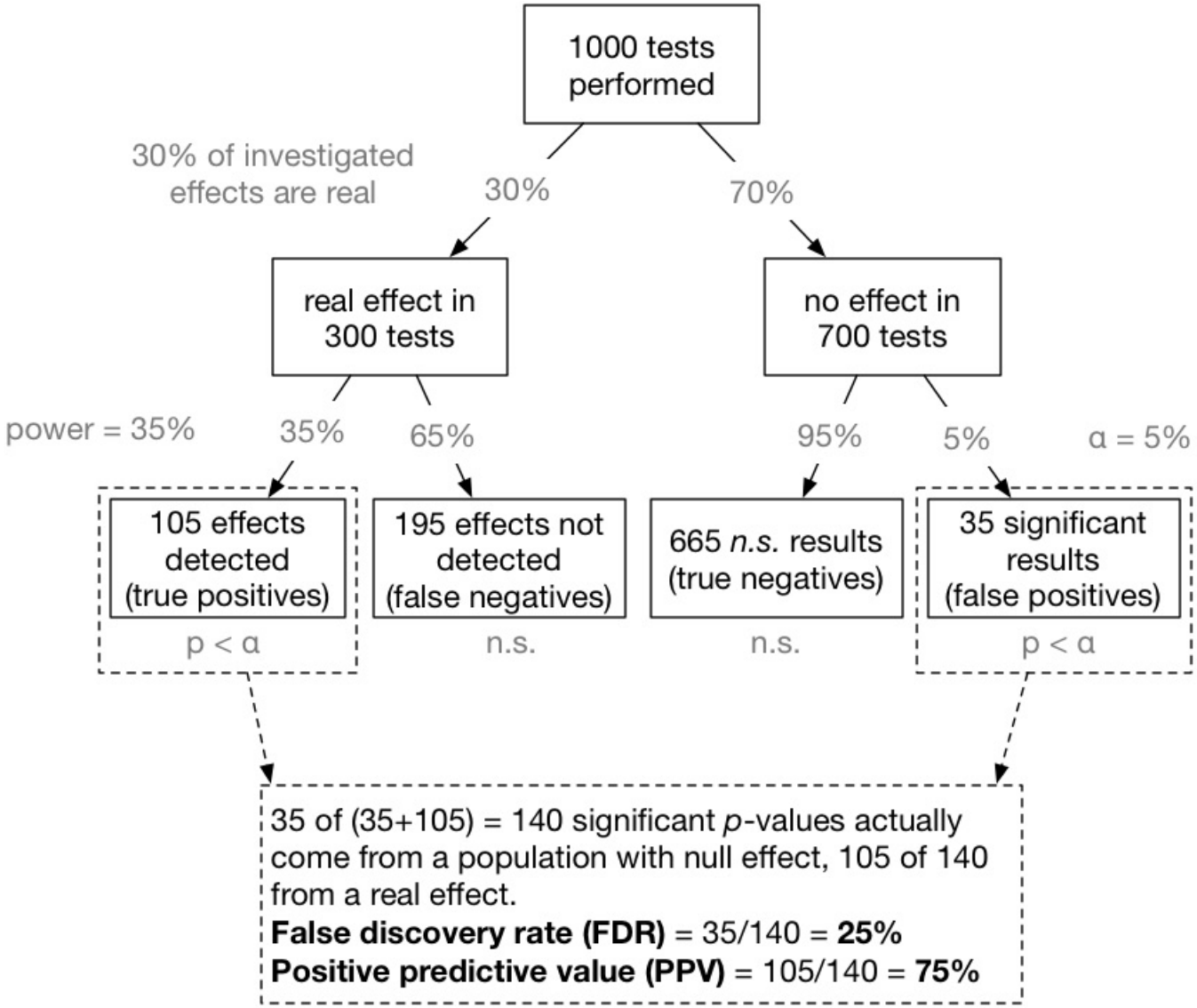
- that H0 is true
- that there is no effect
- that you are a bad scientist
- that you should hide these data (although you probably will not get them published)

p-value > 0.05

it DOES mean

- the data are not surprising, if H0 is true
- H1 might still be true
- not enough statistical power to observe it
- or just bad luck

Why most studies are false?



Interpreting effect sizes

- effects can be statistically significant, but practically insignificant
- with enough data you have statistical power to show even meaningless effects
- two families: Cohen's d and correlations
- Cohen's d: what is the difference between x and y?
- visualization
- correlations: how much the relationship between x and y reduces the error in data?
- visualization

Statistics is tricky

- $P(H|D) \neq P(D|H)$
- p $< \alpha$ does not mean that H1 is true
- p $> \alpha$ does not mean that H1 is false
- if H1 is true, lowering α does not make it more probable to see H1 in data
- if H1 is true, p-values close to α are rather unlikely
- with enough tests, you can "prove" anything
- with enough data, you can detect statistically significant, but practically meaningless effects
- even large effects do not affect everybody

What should happen

- forming hypothesis ahead
- ensuring high statistical power
- controlling false positive error rate
- sound, careful statistical analysis planned ahead
- no tinkering with the results
- reporting effect size and confidence intervals
- publishing even statistically insignificant results
- replicating studies

Conclusion

- statistics is tricky
- our brains do not understand it intuitively
- useful to understand research
- essential to properly do research
- one study does not prove anything
- trouble: low power, high p-value and surprising result
- nothing is certain, but sometimes we are pretty sure

Go and learn more in course