Basics on Hypothesis Tests

APPLIED STATISTICS - FCUP

Rita Gaio
argaio@fc.up.pt

2020

Hypothesis Tests

Hypothesis tests are a basic tool in **Inferential Statistics**

Basic goal: draw (limited. . .) conclusions about a population from sample data

Main contributions

- **Ronald Fisher**: p-values approach for measuring evidence, 1920s. We can use sample data to learn about a population.
- **Jerzy Neyman** & **Egon Pearson**: error rate method (alpha), early 1930s. We cannot learn from individual studies but only from a long series of hypothesis tests.

Textbook publishers, statistics courses, . . . have squished together these two incompatible approaches.

Greenland et al, 2016: *"We have no doubt that the founders of modern statistical testing would be horrified by common treatments of their invention."*

😱 😱 😱

Hypothesis Tests

Framework for parametric hypothesis tests:

- random sample $x_1, \ldots, x_n$ from a r.v. X with probability (density) function $f(x|\theta)$ with unknown θ
- $\theta \in \Theta = \Theta_0 \cup \Theta_1, \quad \Theta_0 \cap \Theta_1 = \emptyset$
- $H_0 : \theta \in \Theta_0$ **null hypothesis** (states the lack of an effect)
  $H_1 : \theta \in \Theta_1$ **alternative hypothesis**
- **Test Statistic**: random variable reflecting the *distance* between sample data and H_0
  - assuming H_0, the test statistic has a known distribution
- **Decision**: define a **rejection region**, **R**, with area equal to a pre-defined accepted error α and compute the test statistic in the sample, $t = T(x_1, \ldots, x_n)$.
  - $t \in R \implies$ reject H_0 and accept H_1, at an α level
  - $t \notin R \implies$ do not reject H_0, at an α level
    (not enough evidence to state H_1)

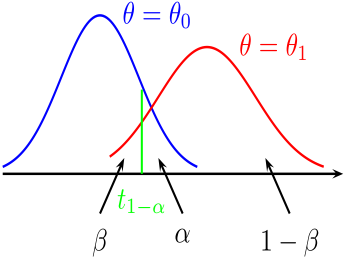
Hypothesis Tests

	H_0 true	H_1 true
Reject H_0	Type I Error	Correct Decision
Do Not Reject H_0	Correct Decision	Type II Error |

Significance Level: $\alpha = P(\text{rej } H_0 \,|\, H_0 \text{ true})$

Power: $P(\text{rej } H_0 \,|\, H_1 \text{ true})$

Want low α and large power but reducing α also decreases the power

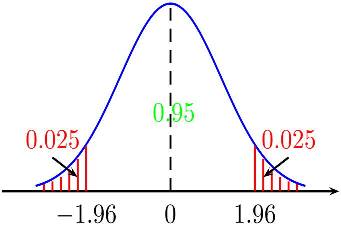


Hypothesis Tests

- Most tests are set up in order to reject H_0 (show statistical evidence for the existence of an effect) thus need to control type I error.
- Fix α (0.05, 0.01, ...) before doing the statistical analysis.
- **Statistically significant** test: whenever we reject H_0.

Hypothesis Tests

Assume that, under H_0, the test statistic Z follows a $N(0,1)$ distribution and the significance level has been fixed at 0.05.



Whenever H_0 is true, $z = Z(x_1, \ldots, x_n)$ falls in the rejection region 5% of the time, considering all possible samples $x_1, \ldots, x_n$. This situation may happen due to random sampling, but we accept to make a 0.05 error by stating that it never occurs whenever H_0 is true. Thus, we reject H_0.

Hypothesis Tests: Neyman and Pearson's view

Neyman and Pearson:

> *no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis.*

i.e., we must abandon our ability to measure evidence, or judge truth, in an individual experiment.

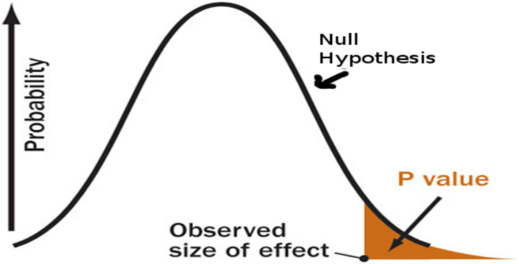
Hypothesis tests:

- were not concerned with which hypothesis (H_0 or H_1) was true or false
- tried to control the overall number of incorrect rejections in the long run.

p-value

p-value: probability of observing a test statistic that is at least as extreme as the one observed in the sample when H_0 is true

p-value: probability that a sample will have an effect at least as extreme as the effect observed in the sample if H_0 is correct.



p-value

Example: $p = 0.225$ -> if H_0 were true, would observe sample effects at least as large as the one observed in the sample about 22.5% of the time, due to random sampling error.

p-value tells how consistent the sample statistics are with H_0:

- high p-values: sample results are consistent with H_0
- low p-values: sample results are not consistent with H_0.

p-value

Wasserstein, R.L., Lazar, N.A. (2016). *The ASA's statement on p-values: context, process, and purpose*, The American Statistician.

- test statistics: normalized difference between data and H_0 model
- p-value varies between 0 and 1
- p-value: continuous measure of compatibility between data and H_0 model
  - $p = 0 \implies$ maximum incompatibility
  - $p = 1 \implies$ maximum compatibility
- p-value: measure of association or effect; not a measure of the size of the effect
- p-value decreases (all the rest kept fixed):
  - as sample size increases
  - as size of the effect increases.

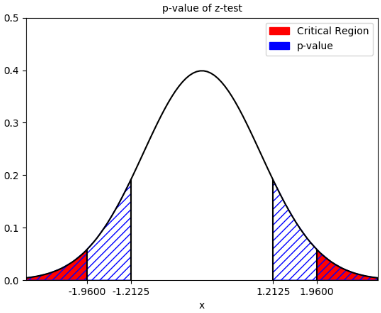
p-value and hypothesis tests

Conflicting Decision:

- p-value $\leq \alpha \implies$ reject H_0 (**statistically significant** test). Our data support the alternative hypothesis.
- p-value $> \alpha \implies$ do not reject H_0

BE CAREFUL when plotting p and α in the same plot!

- false-positive error rate α fixed <u>before</u> the experiment
- p-value computed from a point <u>determined</u> by the data



p-value and α are only superficially similar

p-values ARE NOT error rates

WRONG!

p-**value**: probability of rejecting a null hypothesis that is actually true

WRONG!

Sellke, T., Bayarri, M.J., Berger, J.O. (2001). *Calibration of p Values for Testing Precise Null Hypotheses*, The American Statistician, February 2001, Vol. 55, No. 1

- simulated a large series of hypothesis tests with same H_0
- retained those with p-values ≈ 0.05 $(0.049 \leq p \leq 0.050)$, for which H_0 would thus be rejected
- among those, noted the proportion for which H_0 is true:[1]
  - $P(H_0 \text{ true}) = 0.5 \implies H_0$ was true in 23% of those tests
  - $P(H_0 \text{ true}) = 1/3 \implies H_0$ was true in 12% of those tests
  - ... other initial conditions (see next plot)

[1] These results do not contradict $\alpha = 0.05$ in the sense that α assumes H_0, ie, it assumes $P(H_0 \text{ true}) = 1$

p-values ARE NOT error rates

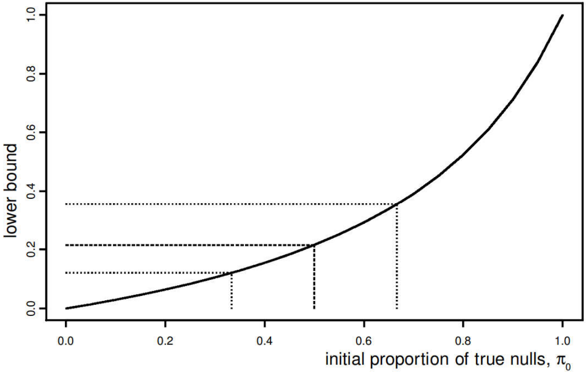


Figure 1. Lower Bound on the Proportion of True Nulls Among Those Tests for Which the p Value is Close to .05.

p-values ARE NOT error rates

Conclusions from the study of Sellke et al:

- a p-value near 0.05 provides a weak evidence against H_0
- however, the lower the p-value, the lower the error rate of incorrectly rejecting H_0.

Example

Goal: show that a new drug has an effect.

Experience: collect two random samples, one taking the new drug and the other taking a placebo. On each individual, read a continuous outcome X.

H_0: $\mu_{\text{drug}} = \mu_{\text{placebo}}$

H_1: $\mu_{\text{drug}} \neq \mu_{\text{placebo}}$

p-value=0.02 means that:

- **Correct**: Assuming the drug has no effect in the population, you'd obtain the sample effect, or larger, in 2% of studies because of random sample error.
- **Incorrect**: There's a 2% chance of making a mistake by rejecting the null hypothesis.

Common wrong interpretations of p-values

WRONG!

- p=0.05 $\implies$ H_0 has a probability of being true of 5%
- p=0.05 $\implies$ there is a 95% or greater chance that H_0 is incorrect

WRONG!

S. T. Goodman, *Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy* (1999). Annals of Internal Medicine, 130 (12).

Indeed, "a p-value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false".

"This logical error reinforces the **mistaken notion** that the data alone can tell us the probability that a hypothesis is true."

Hypothesis Tests

Statistical significance does not necessarily imply practical/scientific relevance; thus

the result from a hypothesis test needs careful interpretation

A p-value from a single study does not provide conclusions with certain error rates but instead adds evidence to that provided by other sources (biological/clinical/scientific/. . .) and other studies.

Hypothesis Tests

Hypothesis tests:

- **robust**: tolerate small deviations from its requirements
- **conservative**: overestimates *p*-value
- **non-conservative**: underestimates *p*-value.

p-values and reproducibility

Replication can increase certainty when findings are reproduced and promote innovation when they are not.

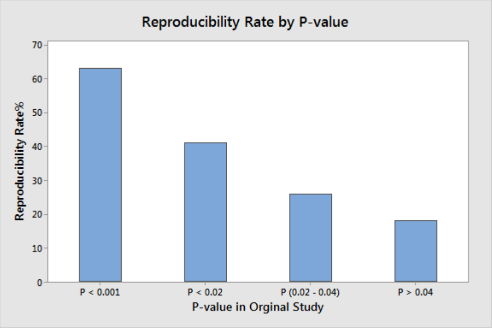
Scientific claims should gain credit because of the replicability of their supporting evidence.

Estimating the reproducibility of psychological science (2015), Science, 349 (6251)

- goal: estimate the reproducibility rate of statistically significant experimental results in psychological studies
- authors are a group of 300 researchers
- they identified 100 psychology studies that had statistically significant findings (97 had p-value$<$ 0.05; 3 had p-value$\approx$ 0.06) and had been published in three top psychology journals. Then they replicated these 100 studies.
- results of each replicate study were compared to those from the corresponding original study.

p-values and reproducibility

- findings: only 36 of the 100 replicate studies were statistically significant -> 36% reproducibility rate!
- lower p-values in the original studies were associated with higher reproducibility rates in the replicate studies.
- 47% of original effect sizes were in the 95% confidence interval of the replication effect size.



p-values and reproducibility

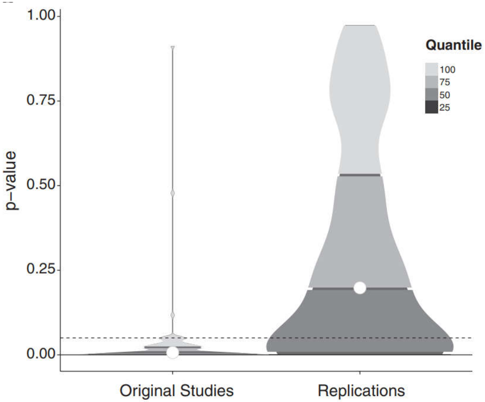


Fig. 1. Density plots of original and replication *P* values and effect sizes.

p-values and reproducibility

- A p-value near 0.05 simply indicates that the result is worth another look; repeated experimentation may be required
- Need to have lower p-values and replicate studies that confirm the initial results before you can safely conclude that an effect exists at the population level.
- Studies with smaller p-values have higher reproducibility rates in follow-up studies.

Guidelines

Solving exercises with hypothesis tests

- name of the test
- identify random variable associated with problem
- make sure requirements are satisfied
- state H_0 and H_1
- identify test statistic and its distribution under H_0; compute/mention test statistic in the sample
- take a conclusion (technical and practical), explaining your answer.

If adequate:

- add an adequate confidence interval to the answer
- compute/mention p-value.