The shortcomings of p-values and the multiple testing bias

Frederick Stegmüller

21. November 2019

Fisher (1925)

- Only reliant on H_0 and the exact p-value
- $p \leq .05$ was proposed as usual threshold, though finally up to the judgement of the experimenter
- H_0 is only demonstrable when experiments rarely give statistically significant results

$\Longrightarrow$ A single significant result is not conclusive until further investigation or replication

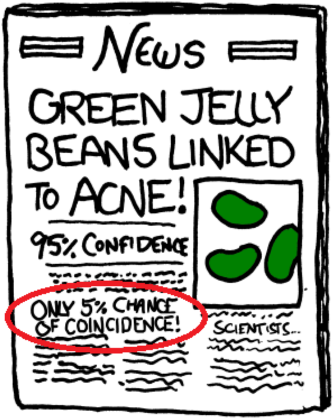
Neyman and Pearson

- Introduced as a formal decision procedure motivated by industrial quality problems
- Goal: minimize false negative rate β and thus maximize power $1 - \beta$ subject to an arbitrary bound α on false positive errors
- Exact p-value is not used as measure of evidence, but to discard H_0 if a critical value is exceeded
- Specific assumptions on H_1 have to be made to minimize β, and an appropriate α has to be chosen
- Designed for repeated testing in the long run, not single experiments

Modern NHST

- H_0 usually predicts no effects, while H_1 is not defined quantitatively
- The p-value is computed and if $p<.05$ H_0 is automatically rejected, while H_1 is accepted and seen as scientific fact
- The exact p-value is then interpreted as a relative measure against H_0

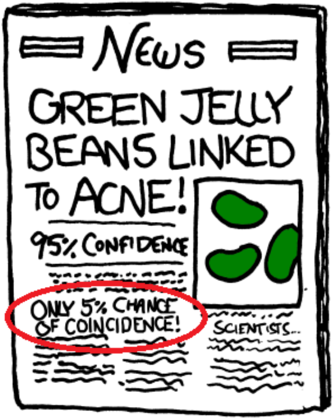
How to interpret p<.05?



source: `https://www.xkcd.com/882/`

- Probability of the null hypothesis given the data? $\rightarrow P(H_0|D)$

How to interpret p<.05?



source: https://www.xkcd.com/882/

- Probability of the data given the hypothesis!
  $\rightarrow P(D|H_0)$

Logical reasoning of rejecting H_0

1. If H_0 is correct, D are highly unlikely
2. D occured

$\Longrightarrow$ $P(D|H_0)$ is highly unlikely, thus we reject H_0 and accept H_1

Logical reasoning of rejecting H_0

1 If H_0 is correct, D are highly unlikely

2 D occured

$\Longrightarrow$ $P(D|H_0)$ is highly unlikely, thus we reject H_0 and accept H_1

1 If a person is an American (H_0), he is unlikely to be a member of congress (D)

2 Trent Kelly is a member of congress

$\Longrightarrow$ Trent Kelly is probably not an American (H_1)

Logical reasoning of rejecting H_0

1. If H_0 is correct, D are highly unlikely
2. D occured

$\Longrightarrow$ $P(D|H_0)$ is highly unlikely, thus we reject H_0 and accept H_1

1. If a person is an American (H_0), he is unlikely to be a member of congress (D)
2. Trent Kelly is a member of congress

$\Longrightarrow$ Trent Kelly is probably not an American (H_1)

- NHST only calculates probabilities concerning H_0, but gives no information on the probability of H_1
- Thus, if multiple H_1s are possible, only vague, non-quantitative statements can be made:
  "The data suggests a significant difference between A and B"

NHST is not suitable for Big Data

- The formula for calculating test statistics depends on the size of the sample (n). Specifically an increase in n leads to a larger test statistic, which then leads to a decrease of the p-value
- Thus, an H_0 expecting a mean of zero, and a large enough sample size can lead to a rejection of H_0 even with miniscule effect sizes

Selective Reporting

- When testing a hypothesis multiple analytical options can result in varying results. This can entice researcher to pick a specific result that is significant, while disregarding other, nonsignificant results
- This can also lead to bias towards a specific hypothesis favored by the researcher

Selective Reporting

Team	Analytic Approach	Distribution	Odds Ratio
10	Multilevel Regression and Logistic Regression	Linear	1.03
1	OLS Regression With Robust Standard Errors, Logistic Regression	Linear	1.18
4	Spearman Correlation	Linear	1.21
14	WLS Regression With Clustered Standard Errors	Linear	1.21
11	Multiple Linear Regression	Linear	1.25
6	Linear Probability Model	Linear	1.28
17	Bayesian Logistic Regression	Logistic	0.96
15	Hierarchical Log-Linear Modeling	Logistic	1.02
18	Hierarchical Bayes Model	Logistic	1.10
31	Logistic Regression	Logistic	1.12
30	Clustered Robust Binomial Logistic Regression	Logistic	1.28
3	Multilevel Logistic Regression Using Bayesian Inference	Logistic	1.31
23	Mixed-Model Logistic Regression	Logistic	1.31
2	Linear Probability Model, Logistic Regression	Logistic	1.34
5	Generalized Linear Mixed Models	Logistic	1.38
24	Multilevel Logistic Regression	Logistic	1.38
28	Mixed-Effects Logistic Regression	Logistic	1.38
32	Generalized Linear Models for Binary Data	Logistic	1.39
8	Negative Binomial Regression With a Log Link	Logistic	1.39
25	Multilevel Logistic Binomial Regression	Logistic	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	Logistic	1.48
7	Dirichlet-Process Bayesian Clustering	Misc	1.71
21	Tobit Regression	Misc	2.88
12	Zero-Inflated Poisson Regression	Poisson	0.89
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	Poisson	1.30
16	Hierarchical Poisson Regression	Poisson	1.32
20	Cross-Classified Multilevel Negative Binomial Model	Poisson	1.40
13	Poisson Multilevel Modeling	Poisson	1.41
27	Poisson Regression	Poisson	2.93

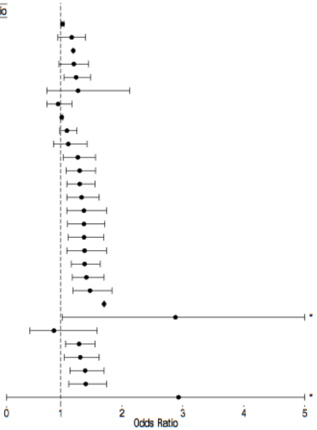


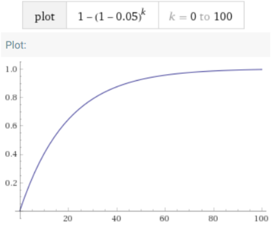
Fig. 3. Point estimates (clustered by analytic approach) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are clustered according to the distribution used in their analyses; within each cluster, the teams are listed in order of the magnitude of the reported effect size, from smallest at the top to largest at the bottom. The asterisks indicate upper bounds that have been truncated to increase the interpretability of the plot (see Fig. 2). OLS = ordinary least squares; WLS = weighted least squares; Misc = miscellaneous.

Publication bias

- Even a true H_0 can, given enough tests, have significant results
- Combined with incentives to only report significant result and disregard insignificant results, any H_0 can be rejected in the long run

Multiple testing bias

- Due to large data sets it is possible to test multiple related hypotheses simultaneously
- Therefore the probability of at least one false positive result (Family-Wise Error Rate) increases with k independent tests if H_0 is true increases: $\alpha_{\text{total}} = 1 - (1 - \alpha)^k$



- This can be accounted for by, among others, the Bonferroni correction for n tests: $p_B = \frac{\alpha}{n}$
- Such corrections, however, decrease the Type I error rate, but increase the Type II error rate

Multiple testing bias

- Alternative to corrections: False Discovery Rate (FDR)
- $Q = \frac{FP}{FP+TP} = \frac{FP}{R}$
- As in research settings the amount of true null effects is not known, only the amount of scnificant and non-significant results, Q can be considered a random variable
- Q cannot be controlled directly, thus define $FDR = E[Q|R > 0] \cdot P(R > 0)$, which can be controled using α and β
- The Family-Wise Error Rate can be defined as $FWER = P(FP \geq 1) = 1 - P(FP = 0)$
- If H_0 is true in all tests, FDR = FWER, while FDR < FWER whith some true H_1

How to do better?

- Change incentives to report all findings, not only significant ones (e.g. pre-registration)
- Publish raw data and analysis scripts
- Don't use arbitrary thresholds to classify into significant ($p=.049$) and non-significant ($p=.051$) findings, rather report p-values as continuous values
- Focus on effect sizes and their uncertainty to form theories, not only on p-values
- Put more importance on reproducing findings and meta-analyses than on the significance of single experiments
- Teach alternative methods and approaches not just NHST

- **Cohen, Jacob.** "The earth is round (p<.05)." What if there were no significance tests? Routledge, 2016. pp. 69-82
- **Ioannidis, John PA.** "What Have We (Not Learnt from Millions of Scientific Papers with P-Values?." The American Statistician, 2019, Vol. 73, pp. 20-25
- **Silberzahn, R., et al.** "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." Advances in Methods and Practices in Psychological Science, 2018, Vol. 1(3), pp. 337-356
- **Szucs, Denes, and Ioannidis, John.** "When null hypothesis significance testing is unsuitable for research: a reassessment." Frontiers in human neuroscience, 2017, Vol. 11, Article 390
- **Wasserstein, Ronald L., Schirm, Allen L., and Lazar, Nicole A.** "Moving to a World Beyond "p<0.05"." The American Statistician, 2019, Vol 73, pp. 1-19