P-Value Abuse

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Nirogacestat, a γ-Secretase Inhibitor for Desmoid Tumors

Mrinal Gounder, M.D., Ravin Ratan, M.D., Thierry Alcindor, M.D., Patrick Schöffski, M.D., M.P.H., Winette T. van der Graaf, M.D., Ph.D., Breelyn A. Wilky, M.D., Richard F. Riedel, M.D., Allison Lim, Pharm.D., L. Mary Smith, Ph.D., Stephanie Moody, M.S., Steven Attia, D.O., Sant Chawla, M.D., et al.

across prespecified subgroups. The percentage of patients who had an objective response was significantly higher with nirogacestat than with placebo (41% vs. 8%; P<0.001)

P values in science

Often a requisite for publication

Should not be the only criterion, because they

- Do not imply causal effects
- Do not imply practical significance

Also encourages publication bias / p-hacking

Randomized control trial

Goal: Evaluate cure rate of two expensive drugs with side effects

Drug 1:

Control group: 30 out of 100

Treatment group: 52 out 100

Drug 2:

Control group: 30,000 out of 100,000

Treatment group: 30,650 out of 100,000

Two-sample z test

Null hypothesis: All data are i.i.d. Bernoulli with cure rate θ_{null}

Test statistic: Difference in cure rate between treatment and control groups

Under null hypothesis, Gaussian with mean 0 and variance

$$\sigma_{\text{null}}^2 := \theta_{\text{null}}(1 - \theta_{\text{null}}) \left(\frac{1}{n_{\text{treatment}}} + \frac{1}{n_{\text{control}}} \right)$$

$$\text{pv}(t_{\text{data}}) = \mathrm{P}\left(\tilde{t}_{\text{null}} \geq t_{\text{data}}\right)$$

Drug 1: $t_{\text{data}} = 0.220$ $\sigma_{\text{null}} = 6.96 \cdot 10^{-2}$ $\text{pv}(t_{\text{data}}) = 7.8 \cdot 10^{-4}$

Drug 2: $t_{\text{data}} = 0.007$ $\sigma_{\text{null}} = 2.06 \cdot 10^{-3}$ $\text{pv}(t_{\text{data}}) = 7.8 \cdot 10^{-4}$

What does this mean?

Both results are equally unlikely under null hypothesis

We're pretty sure both drugs increase cure rate

Is this all we care about? No!

How can we quantify by how much they increase it?

Confidence interval for difference in cure rate

Difference in cure rate

True control cure rate: θ_C

Number of cured control subjects $\tilde{k}_C$:

Binomial with parameters n_C and θ_C

$\approx$ Gaussian with mean $n_C \theta_C$ and variance $n_C \theta_C (1 - \theta_C)$

Observed control cure rate $\tilde{k}_C / n_C$:

$\approx$ Gaussian with mean θ_C and variance $\theta_C (1 - \theta_C) / n_C$

Difference in cure rate

True treatment cure rate: θ_T

Observed treatment cure rate: $\tilde{k}_T / n_T$:

$\approx$ Gaussian with mean θ_T and variance $\theta_T(1-\theta_T)/n_T$

Difference: $\tilde{k}_T / n_T - \tilde{k}_C / n_C$:

$\approx$ Gaussian with mean $\theta_T - \theta_C$ and variance

$$\sigma^2 := \frac{\theta_T(1-\theta_T)}{n_T} + \frac{\theta_C(1-\theta_C)}{n_C}$$

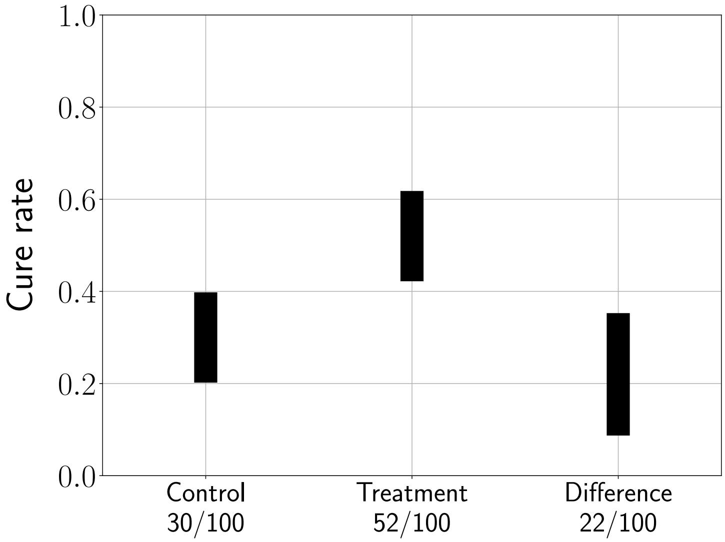
Confidence interval for a Gaussian

Let $\tilde{a}$ be Gaussian with mean μ and variance σ^2

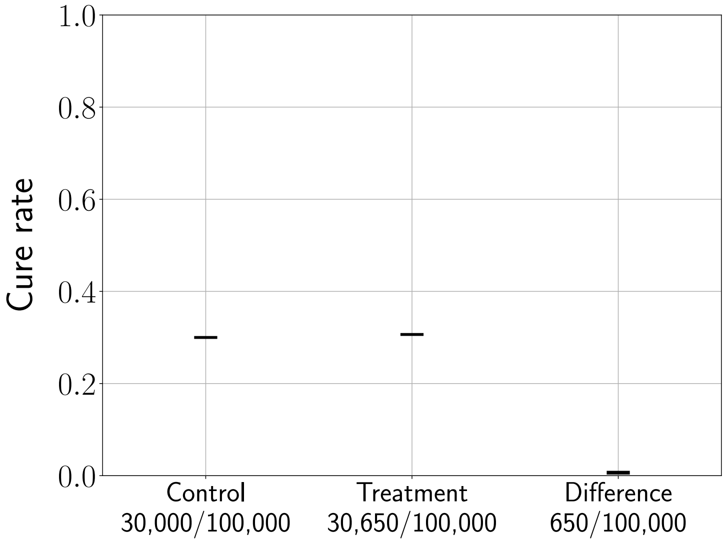
$$\widetilde{\mathcal{I}}_{1-\alpha} := [\tilde{a} - c_\alpha \sigma, \tilde{a} + c_\alpha \sigma] \qquad c_\alpha := F_{\tilde{z}}^{-1}\left(1 - \frac{\alpha}{2}\right)$$

$$\widetilde{\mathcal{I}}_{0.95} := [\tilde{a} - 1.96\sigma, \tilde{a} + 1.96\sigma]$$

Drug 1: $[8.71\%, 35.2\%]$



Drug 2: $[0.25\%, 1.05\%]$



Statistical vs practical significance

In large-scale trials, tiny differences can be statistically significant

COVID-19 vaccine

43,448 patients randomly divided into

- Treatment group of 21,720 patients: 8 cases (0.037%)
- Control group of 21,728 patients: 162 (0.746%)

$$\mathrm{pv}(t_{\mathrm{data}}) < 10^{-23}$$

Fictitious vaccine trial

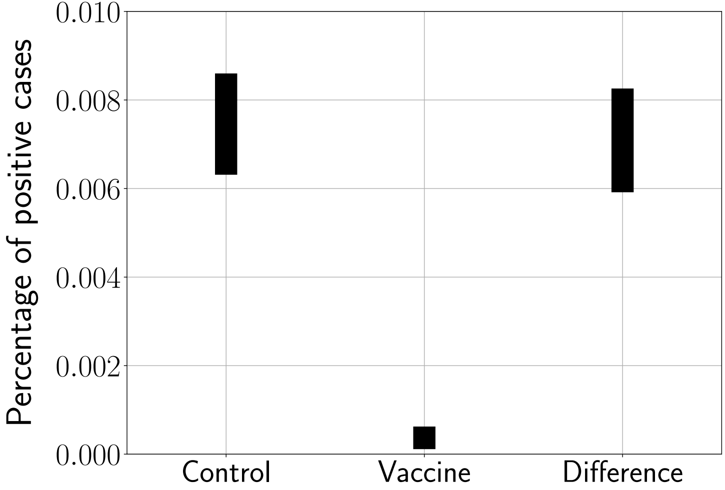
43,448 patients randomly divided into

- Treatment group of 21,720 patients: 120 cases (0.552%)
- Control group of 21,728 patients: 162 (0.746%)

$$\mathrm{pv}(t_{\mathrm{data}}) = 0.006$$

Ratio of positive cases is 3/4, not practically significant! (Real data: 1/20)

Actual vaccine trial



Obama's presidential campaign

Options: Image or video on website

Metric: Sign-up rate

- Images: 14,016 out of 155,280
- Videos: 10,337 out of 155,102

$$\mathrm{pv}(t_{\mathrm{data}}) < 10^{-80}$$

Fictitious experiment

Options: Image or video on website

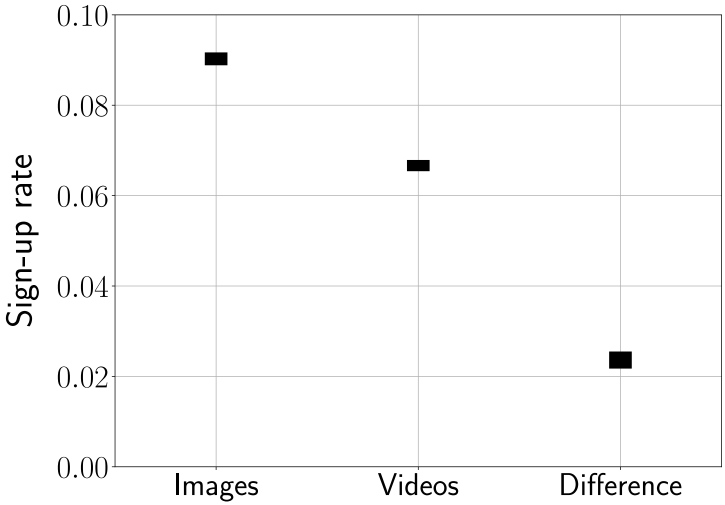
Metric: Sign-up rate

- Images: 14,016 out of 155,280
- Videos: 13,650 out of 155,102

$$\text{pv}(t_{\text{data}}) < 0.027$$

Difference in sign-up rate is 0.0002, not practically significant!!

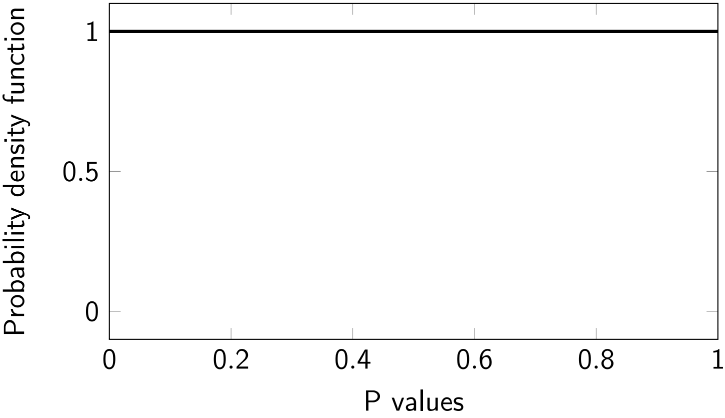
Actual experiment



Pizza and COVID-19

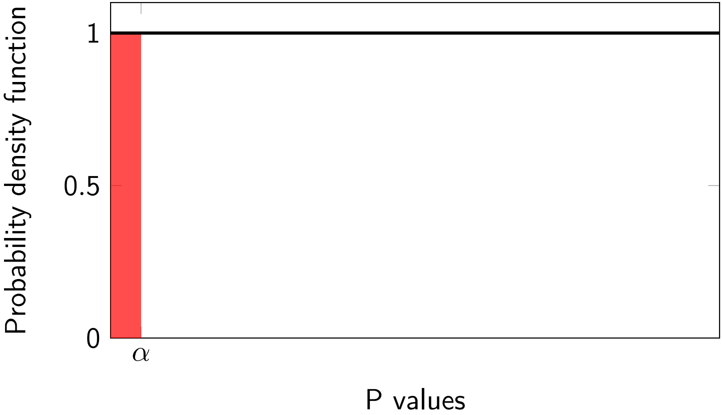
100 studies to determine whether pizza cures COVID-19

P-value distribution? (Test statistic is continuous)



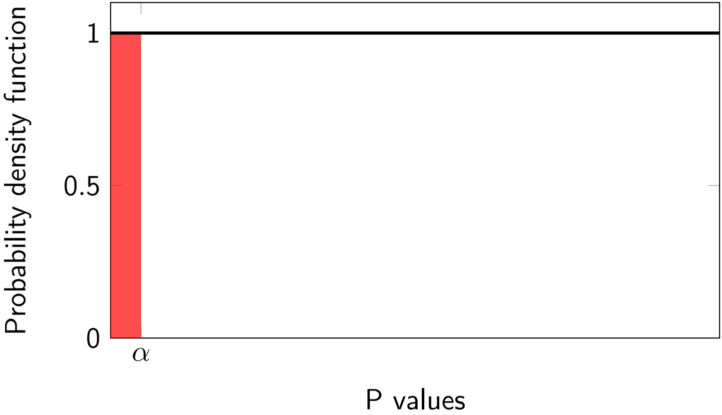
Significance level $\alpha := 0.05$

Probability of a single false positive? 0.05



Significance level $\alpha := 0.05$

Under null hypothesis, fraction of false positives among many tests? 0.05



Pizza and COVID-19

100 studies to determine whether pizza cures COVID-19

≈ 95 true negatives

≈ 5 false positives

If all results are published no problem

Unfortunately, much easier to publish if result is statistically significant!

Publication bias: You only hear about the false positives!

Food additives

We test many food additives on mice

One of them yields small p value

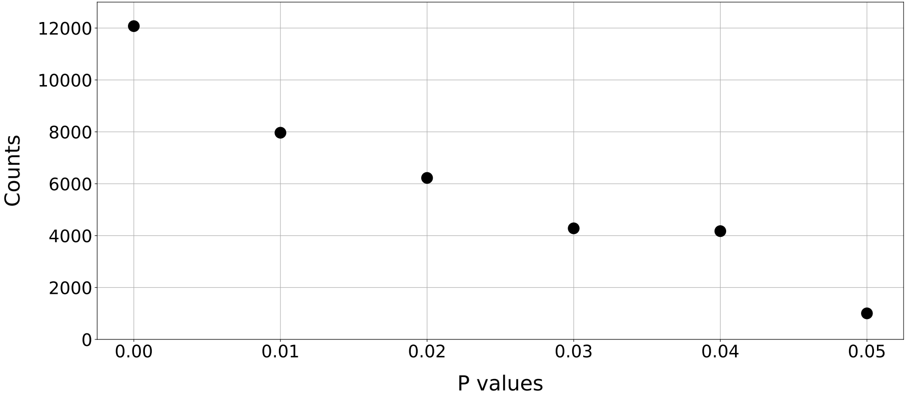
But not significant after Bonferroni's correction

Two options

1. Gather additional data
2. Publish result (p hacking!)

Does p hacking occur in practice?

Distribution of p values in PubMed[1]



[1] Head, M. L, Holman, L., Lanfear, R., Kahn, A. T, and Jennions, M. D. *The extent and consequences of p-hacking in science*. PLoS biology

What have we learned

P values are very useful, but should not be the only criterion to evaluate a finding!

- They do not imply practical significance
- Publication bias / p-hacking