Adjusting the *p*-values' role in Biomedical Research

Horacio Gómez-Acevedo
Department of Biomedical Informatics
University of Arkansas for Medical Sciences

September 16, 2021

What is a *p*-value?

Ingredients:

- Data (random sample)
- Statistic (a special type of function)
- A value (called the level of significance)
- A (null) hypothesis H_0

Definition

The *p*-value is the observed tail probability of a statistic being at least as extreme as the particular observed value when H_0 is true.

What is a p-value?(cont)

Analog in Criminal trials
If the defendant is innocent, what is the chance that we would observe such extreme criminal evidence?
Source

The problem of statistical significance

In the mind of hypothetical researcher:

> *I used one-way ANOVA to differentiate the femur BMD between wildtype and two genotypes and I got p-value* < 0.049*, thus the results are* ***statistically significant****” ... (champagne is in order!!)*
>
> *However, I checked body weight of those genotypes at 6 months and I got p-value* $= 0.051$*. I cannot reproduce what I saw last time!... (throw the experiment away and never talk about it!!)*

What is the problem with that?

The p-value and the threshold (typically 0.05) should NEVER (EVER) used as a test for no association or no difference.

In other words

> ***STOP CATEGORIZING BASED ONLY ON*** p-***VALUES***

The phrase **statistically significant** should be discontinued because it supports false assumptions of difference or lack thereof.

It is not a new problem!

It has been on statistical circles (and research) for more than a decade. The American Statistical Association had an special issue in 2018 about this very topic on the journal **The American Statistician**.

> *When I see articles with lots of significance test, I say that the statisticians are p-ing on the research*
> *Herman Friedman*

What does the ASA say?

In 2016, the American Statistical Association released an statement about this:
- *p*-values can indicate how incompatible the data are with a specified statistical model.
- *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

What does the ASA say? (cont)

- Scientific conclusions and business or policy decisions should not be based only on whether a *p*-values passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Some important points

- No *p*-value can reveal the plausibility, presence, truth, or importance of an association or effect.
- A label of statistical significance does not mean or imply that an association or effect is highly probable, real, true, or important.
- Nor does a label of statistical nonsignificance lead to the association or effect being improbable, absent, false, or unimportant

  Significant" and "not significant" is not itself statistically significant.
  Gelman and Stern (2006)

We are affecting the scientific process

Jessica Utts (ASA former president)
"Over time it appears that p-value become a gatekeeper for whether work is publishable ... this apparent editorial bias leads to the *file-drawer effect* in which research with statistically significant outcomes are more likely to get published, while other work that might well be just as important scientifically is never seen print."
The end result will bias the very research endeavour that we are involved in..

Sort of good news?

Wasserstein et al (2018) point out that

- There is not a solution that majestically replaces the outsized role that statistical significance has come to play.
- The statistical community has not yet converged on a simple paradigm for the use of statistical inference in scientific research.

Sort of good news?

Recommendations, use the **ATOM** methodology

- **A**ccept uncertainty.
- Be **T**houghtful,
- **O**pen, and
- **M**odest

Accept uncertainty

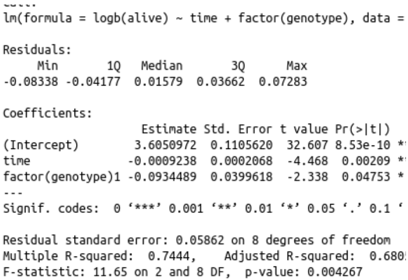
- With every point estimate we add a measure of its uncertainty (e.g., standard error, interval estimate).
- Report and interpret point and interval estimate.
- Start thinking of confidence intervals as ”compatibility intervals” which use *p*-values to show the effect sizes that are most compatible with the data under the given model.

Be Thoughtful

Answer some (or all) of these questions

- What are the practical implications of the estimate?
- How precise is the estimate?
- Is the model correctly specified?
- Are the modeling assumptions understood?
- Are the assumptions valid?
- Do the key results hold up when other modeling choices are made?

Be Thoughtful (cont.)

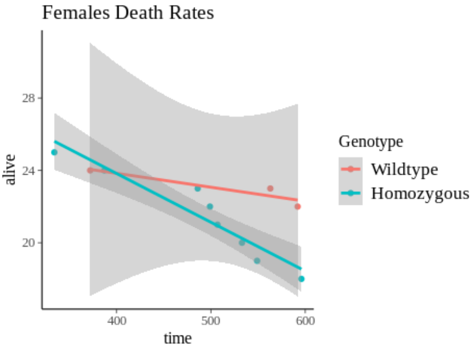


```
Call:
lm(formula = logb(alive) ~ time + factor(genotype), data =

Residuals:
     Min       1Q   Median       3Q      Max
-0.08338 -0.04177  0.01579  0.03662  0.07283

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.6050972  0.1105620  32.607 8.53e-10 **
time             -0.0009238  0.0002068  -4.468  0.00209 **
factor(genotype)1 -0.0934489  0.0399618  -2.338  0.04753 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.05862 on 8 degrees of freedom
Multiple R-squared:  0.7444,    Adjusted R-squared:  0.680
F-statistic: 11.65 on 2 and 8 DF,  p-value: 0.004267
```



Be Thoughtful (and fancy)

There are multiple approaches to complement the p-values.

- Greenland et al (2019) suggest the use s-values (Shannon information $s = -\log_2(p)$) that moves users away from probability misinterpretation.
- Second generation p-values (SGPV) from Blume et al. They claim that it lowers false discovery rates and more likely it would be reproducible in future studies.

Be Open

- Public pre-registration of methods
- Transparency
- Completeness in reporting

Personal decision making is part of statistical data analysis, but it should be openly disclosed, so other researchers can execute maninful alternative analyses.

Be Modest

- There is not a ”true statistical model” underlying every problem.
- *p*-values, confidence intervals and other statistical measures are all uncertain.
- The nexus of opennes and modesty is to report everything while at the same time not concluding anything from a single study with unwarranted certainty.
- Encourage others to reproduce your work.

Change is slow

We can start with replacing

- *p*-values with an equality (not inequality)
- if you obtain $p = 0.03$" avoid using the phrase **statistically significant**
- if you obtain $p = 0.25$" avoid using the phrase **was not statistically significant** just report the value.
- compare groups and studies directly by showing *p*-values and interval estimates for their differences, for instance *p*-values and confidence interval for the different for sex-specific associations.
- never use the phrase **approaching to significance**

Book Example

Sixty-six women with osteoporosis were alternately assigned to one of three treatment groups: Group 1 ($n = 22$), group 2 ($n = 22$), and controls ($n = 22$). After 6 weeks, the change in BMD from baseline was measured. Analysis with one-way ANOVA indicated a statistically significant difference between the groups ($F_{2,63} = 61.07; P < 0.0001$) had a $p = 1 \times 10^{-4}$ ($F_{2,63} = 61.07$).

Book Example (cont.)

Further analysis with Tukey's pair-wise comparison procedure to control for multiple testing revealed that the mean change ($\pm$SD) of group 2 (1.6 $g/cm^2 \pm 0.2$) was significantly greater than of group 1 (1.1 $g/cm^2 \pm 0.2$) and that of the controls (1.0 $g/cm^2 \pm 0.2$) with an overall alpha level of 0.05. group 2 has a higher mean (1.6 g/cm^2), than group 1 (1.1 g/cm^2) and controls (1.0g/cm^2). The 90% confidence intervals for groups 2, 1 and controls are $(1.5, 1.8)$, $(1.05, 1.15)$, and $(0.85, 1.15)$, respectively. From Land and Secic: How to report statistics in medicine (2005)

Final thoughts

- Abuse of (click and play) statistical software
- It obscures reproducibility
- Some of the warnings of modeling are hidden
- New methodologies are not implemented.

When you see a *p*-value

Recall ATOM (ant)