Homework 3: P-values

Joe Brew

January 26, 2015

joebrew@gmail.com
UFID: 0402-8902
+001 352 318 4553

Contents

Homework 3: P-values

1. Rothman's six misconceptions

In his "Six Persistent Research Misconceptions", Kenneth Rothman outlines six beliefs that persist *despite* substantial evidence that these ways of thinking are flawed.[1] I outline these six misconceptions, with both a definition and personal reflection.

1. **There is a hierarchy of study designs**: Rothman claims that "absolute proof" can never be demonstrated, and backs this up with the assertion that variability in the results of randomized trials is evidence *against* perfection. He goes on to explain systematic error, non-adherence, and measurement error. Even among non-RCTs, the supposed hierarchy cohort > case-control) is a myth, as both have strengths and weaknesses. My personal opinion is that Rothman is correct, but fails to acknowledge that RCTs *do* have many advantages which observational studies can never achieve.

2. **Representativeness is required for generalization**: Rothman distinguishes "statistical" and "scientific" generalizability. For the latter, we simply need to make a case, acknowledging the population from which we are generalizing. I agree with his argument, especially in light of his qualification that "representativeness" should be required as a function of the plausability of differential effects.

3. **If an interaction term isn't 'significant' in a regression output then biological interaction doesn't exist**: In statistical outputs, the interpretability can be rendered nearly useless by the units used (a hodge-podge of time, quantity, quality, mass, etc.). I agree with Rothman that how we measure and categorize can radically affect whether a truly biologic interaction is detected statistically.

4. **Percentile boundaries are appropriate for binning**: Rothman points out two shortcomings in the use of percentiles: (1) that these may or may not correspond to biologically important cut-offs and (2) that study-specific quantiles make cross-study interpretability impossible. I agree with Rothman, and think that his suggestion to use progressively larger binning algorithms is helpful to both making results interpretable and biologically sound.

5. **One should always quantify uncertainty in multiple regression**: Rothman highlights the fact that multiple regression increases the likelihood of the detection of a spurious result. According to Rothman, we should worry less about type 1 error. I appreciate Rothman's suggestion that we instead use a more Bayesian approach, but would have appreciated slightly more detail on *why* this is a superior method.

6. **Significance testing is important for interpretation of data**: To Rothman, significance tests are simply a "degraded" mirror of p-values. We should instead be estimating effects, and not performing so much "statistical testing." Rothman is correct to point out that a confidence interval contains *all* of the information needed to understand the result: P-values and significance delimitations are entirely unnecessary.

I appreciated Rothman's article, particularly the succinctness with which he lays out (and provides examples for) his 6 criticisms of the field.

2. Three definitions of P-values and 95% confidence intervals

In Andrew Gelman's "P Values and Statistical Practice", the author describes the P-value as "a measure of discrepancy of the fit of a model or 'null hypothesis' H to data y." [2]

In a 2013 article, Schuemie et al make the case that the typically-held definition that "$p < 0.5$" means "there is only a 5% probability that the observed effect would be seen by chance when in reality there is no effect" is entirely incorrect, particularly in the case of observational studies. [3] They propose a "calibrated p-value" which is a p-value calculated as such: "the error distribution resulting from this bias (which does not depend on sample size) can be added to the random error distribution (which is based on sample size)."

In a paper on Bayesian statistics, Sander Greenland describes how P-values differ in regression outputs as a function of the type of statistical thinking employed (frequentist vs. Bayesian). [4] Though (s)he does not provide a direct definition, Greenland describes the interpretation of P-values as such: "A small P-value indicates incompatibility of the prior with the likelihood, which could arise from faulty prior information, faulty actual data, a faulty likelihood model, or some combination. A large P-value, however, does not mean the prior and likelihood are compatible, let alone correct; at best one can only say that the diagnostic detected no problem."

3. Goodman's P-value fallacy

Two sentence summary: The "fallacy" is the "idea that a single number can capture both the long-run outcomes of an experiment and the evidential meaning of a single result." [5] In other words, Goodman posits that trying to combine both a statistical and scientific indicator (as well as "long run" and "short run" indicator) of quality/significance/remarkability is both impractical and inaccurate, echoing Rothman's sentiments.[1]

Do confidence itnervals solve the fallacy? No - though "they push us away from the automaticity of P-values and hypothesis tests by promoting a consideration of the size of the observed effect", they also "embody, albeit in a subtler form, many of the same problems that afflict current methods, the most important being that they offer no mechanism to unite external evidence with that provided by an experiment" (I would say more, but we're limited to one sentence!)

References

[1] Kenneth J. Rothman. Six persistent research misconceptions. *J GEN INTERN MED*, 29(7): 1060–1064, jan 2014. doi: 10.1007/s11606-013-2755-z. URL `http://dx.doi.org/10.1007/s11606-013-2755-z`.

[2] Andrew Gelman. P values and statistical practice. *Epidemiology*, 24(1):69–72, jan 2013. doi: 10.1097/ede.0b013e31827886f7. URL `http://dx.doi.org/10.1097/EDE.0b013e31827886f7`.

[3] Martijn J. Schuemie, Patrick B. Ryan, William DuMouchel, Marc A. Suchard, and David Madigan. Interpreting observational studies: why empirical calibration is needed to correct p -values. *Statist. Med.*, 33(2):209–218, jul 2013. doi: 10.1002/sim.5925. URL `http://dx.doi.org/10.1002/sim.5925`.

[4] S. Greenland. Bayesian perspectives for epidemiological research. II. regression analysis. *International Journal of Epidemiology*, 36(1):195–202, feb 2007. doi: 10.1093/ije/dyl289. URL `http://dx.doi.org/10.1093/ije/dyl289`.

[5] Steven N. Goodman. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12):995, jun 1999. doi: 10.7326/0003-4819-130-12-199906150-00008. URL `http://dx.doi.org/10.7326/0003-4819-130-12-199906150-00008`.

Details

Full code at https://github.com/joebrew/uf/tree/master/phc7000.

This report was generated on January 26, 2015. The author used R version 3.1.1 (2014-07-10) (Sock it to Me) on a mingw32 OS.

The analysis in this report was written in the R programming language, and the report production was programmed in LaTeX using Sweave.