Who's gotta p?

Tripping up statistics in the garden of forking paths

Phillip M. Alday
(UniSA)

19 October 2016

What's a *p*-value?

Let's get the hard part done up front

Definition

The p-value is the probability under the null hypothesis of a result at least as extreme as the observed result.

Let's get the hard part done up front

Definition

The p-value is the probability under the null hypothesis of a result at least as extreme as the observed result.

A whole other can of worms: NHST

- the p-value is useless as soon as we reject the null hypothesis
- the p-value is uniformly distributed under the null hypothesis (Murdoch, Tsai, and Adcock 2008)

But I'm not here to bash the p-value (too much)

For better or worse, there are a lot of p-values out there and we have to deal with them (cf. Gelman 2013).

But I'm not here to bash the p-value (too much)

For better or worse, there are a lot of p-values out there and we have to deal with them (cf. Gelman 2013).

So what do we have to watch out for?

Assume makes an ass out of u and me

What happens if you violate testing assumptions?

What happens if you violate testing assumptions?

Not much (for many methods).

What happens if you violate testing assumptions?

Not much (for many methods).

- At least in terms of your point estimates

What happens if you violate testing assumptions?

Not much (for many methods).

- At least in terms of your point estimates
- Well, with the exception of outliers and the like (try robust statistics instead, cf. Wilcox 2010)

What happens if you violate testing assumptions?

Not much (for many methods).

- At least in terms of your point estimates
- Well, with the exception of outliers and the like (try robust statistics instead, cf. Wilcox 2010)
- And lack of power may be a bigger issue anyway (Button et al. 2013; Gelman 2015b)

What happens if you violate testing assumptions?

Not much (for many methods).

- At least in terms of your point estimates
- Well, with the exception of outliers and the like (try robust statistics instead, cf. Wilcox 2010)
- And lack of power may be a bigger issue anyway (Button et al. 2013; Gelman 2015b)

What happens if you violate testing assumptions?

Not much (for many methods).

- At least in terms of your point estimates
- Well, with the exception of outliers and the like (try robust statistics instead, cf. Wilcox 2010)
- And lack of power may be a bigger issue anyway (Button et al. 2013; Gelman 2015b)

A whole lot in terms errors.

- Error estimates no longer guaranteed to be correct

What happens if you violate testing assumptions?

Not much (for many methods).

- At least in terms of your point estimates
- Well, with the exception of outliers and the like (try robust statistics instead, cf. Wilcox 2010)
- And lack of power may be a bigger issue anyway (Button et al. 2013; Gelman 2015b)

A whole lot in terms errors.

- Error estimates no longer guaranteed to be correct
- *p*-values deeply intertwined with error estimates

What happens if you violate testing assumptions?

Not much (for many methods).

- At least in terms of your point estimates
- Well, with the exception of outliers and the like (try robust statistics instead, cf. Wilcox 2010)
- And lack of power may be a bigger issue anyway (Button et al. 2013; Gelman 2015b)

A whole lot in terms errors.

- Error estimates no longer guaranteed to be correct
- *p*-values deeply intertwined with error estimates
- *p*-values now effectively garbage

How are *p*-values actually computed?

0. Under the null hypothesis, test statistics have a known distribution
   - either analytic: t, F, χ^2, etc.
   - or empirically determined: bootstrapping and other resampling methods
1. Compute the test statistic
   - t-test: mean divided by standard error
   - F-test: between vs. within group variance
   - degrees of freedom compensate for small samples
2. Compare it to the reference distribution
   - Old way: use a table and find out what your p-value is less than
   - New way: have the computer figure out a precise value

Uh oh . . .

When we violate testing assumptions

- the error term is wrong
- and so the test statistic is no longer guaranteed to follow the assumed distribution.
- (even resamping techniques such as the bootstrap have assumptions, albeit different ones)

Uh oh . . .

When we violate testing assumptions

- the error term is wrong
- and so the test statistic is no longer guaranteed to follow the assumed distribution.
- (even resamping techniques such as the bootstrap have assumptions, albeit different ones)

BAD NEWS BEARS

- We no longer know the proper reference distribution.
- And so we know longer know what null hypothesis (whether or not meaningful!) the p-value corresponds to.

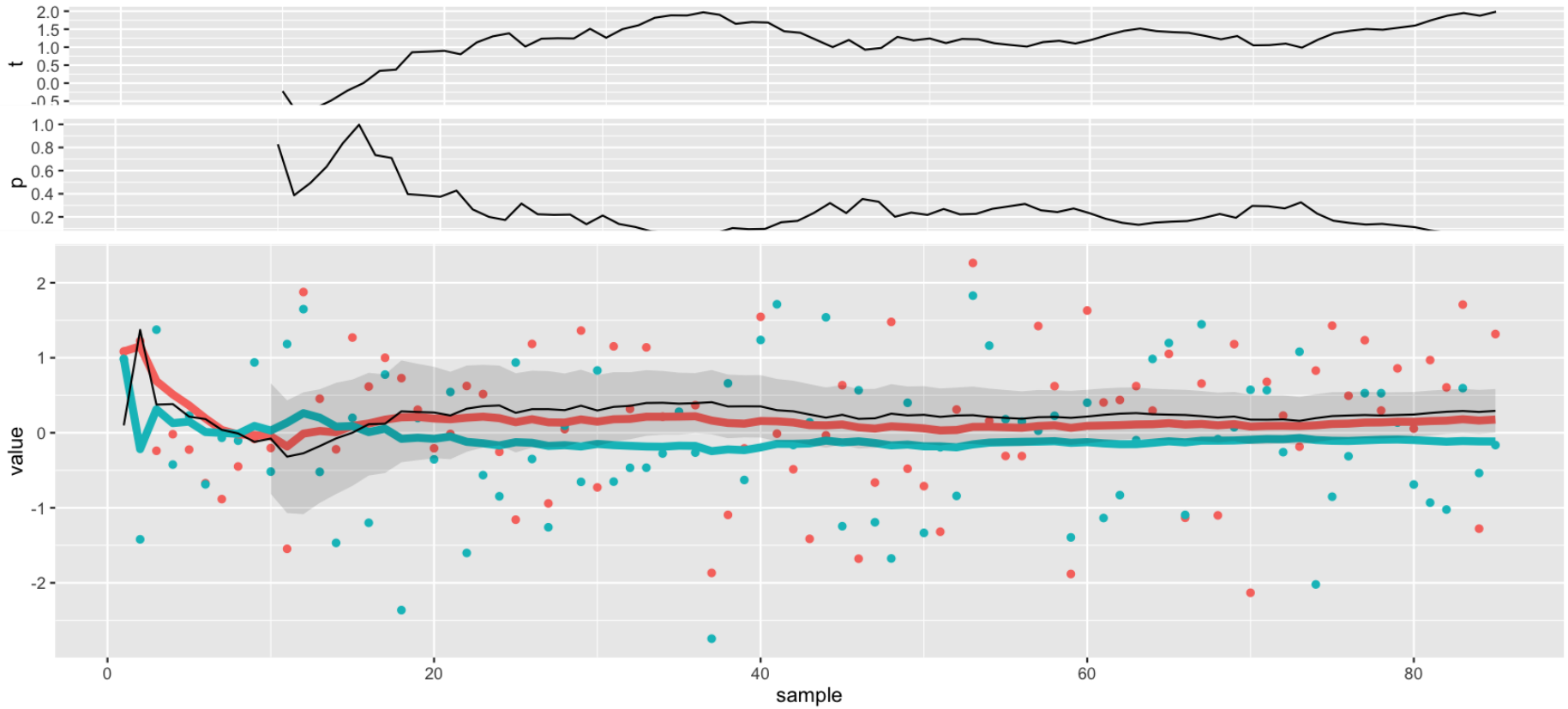
What I really want to say is

Don't use t / F / traditional ANOVA for accuracy data (Jaeger 2008; Allefeld, Görgen, and Haynes 2016).

Don't use inferential statistics for comparing non-randomly sampled, closed populations such as stimuli or experimental groups (Sassenhagen and Alday 2016).

Optional Stopping

We're not so different



try it yourself at `https://tinyurl.com/optionalstopping`

Optional stopping is a great way to guarantee a significant yet meaningless result

If you're going to play the significance game, play by the rules.

- Set your α-threshold ahead of time
  - no "trending towards significance"
  - no "marginal signficance"
- Determine your stopping rule ahead of time
  - power analysis
  - rules of thumb
  - "until we run out of money"
- But don't base it on peeking at *p*-values
- Only claim the significance you test (the difference between significant and not significant is not itself significant; cf. Gelman and Stern 2006; Nieuwenhuis, Forstmann, and Wagenmakers 2011)

Never forget

The *p*-value is a random variable under the null hypothesis.

- *p*-hacking and "fishing expeditions" will always succeed
- And if you're not careful, you'll find yourself far out at sea

Welcome to the garden of forking paths

Blurring the distinction exploratory and confirmatory makes interpretable *p*-values hard to find

- There is (almost) always a significant result lurking in any given dataset just by chance
  - "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011)
  - "garden of forking paths" (Gelman and Loken 2013)

Blurring the distinction exploratory and confirmatory makes interpretable *p*-values hard to find

- There is (almost) always a significant result lurking in any given dataset just by chance
  - "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011)
  - "garden of forking paths" (Gelman and Loken 2013)
- We needn't care about significance; we must care about *scientific truth*

Blurring the distinction exploratory and confirmatory makes interpretable *p*-values hard to find

- There is (almost) always a significant result lurking in any given dataset just by chance
  - "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011)
  - "garden of forking paths" (Gelman and Loken 2013)
- We needn't care about significance; we must care about *scientific truth*
- Focus on accurately estimating your effects, making falsifiable predictions and then testing those predictions, and the rest will follow (cf. Wagenmakers et al. 2012; Kruschke 2013; Cumming 2014; Gelman 2015a; Button 2016; Yarkoni and Westfall 2016)

Blurring the distinction exploratory and confirmatory makes interpretable *p*-values hard to find

- There is (almost) always a significant result lurking in any given dataset just by chance
  - "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011)
  - "garden of forking paths" (Gelman and Loken 2013)
- We needn't care about significance; we must care about *scientific truth*
- Focus on accurately estimating your effects, making falsifiable predictions and then testing those predictions, and the rest will follow (cf. Wagenmakers et al. 2012; Kruschke 2013; Cumming 2014; Gelman 2015a; Button 2016; Yarkoni and Westfall 2016)
- And if your results are falsified, then we've still learned something

Blurring the distinction exploratory and confirmatory makes interpretable *p*-values hard to find

- There is (almost) always a significant result lurking in any given dataset just by chance
  - "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011)
  - "garden of forking paths" (Gelman and Loken 2013)
- We needn't care about significance; we must care about *scientific truth*
- Focus on accurately estimating your effects, making falsifiable predictions and then testing those predictions, and the rest will follow (cf. Wagenmakers et al. 2012; Kruschke 2013; Cumming 2014; Gelman 2015a; Button 2016; Yarkoni and Westfall 2016)
- And if your results are falsified, then we've still learned something

Blurring the distinction exploratory and confirmatory makes interpretable *p*-values hard to find

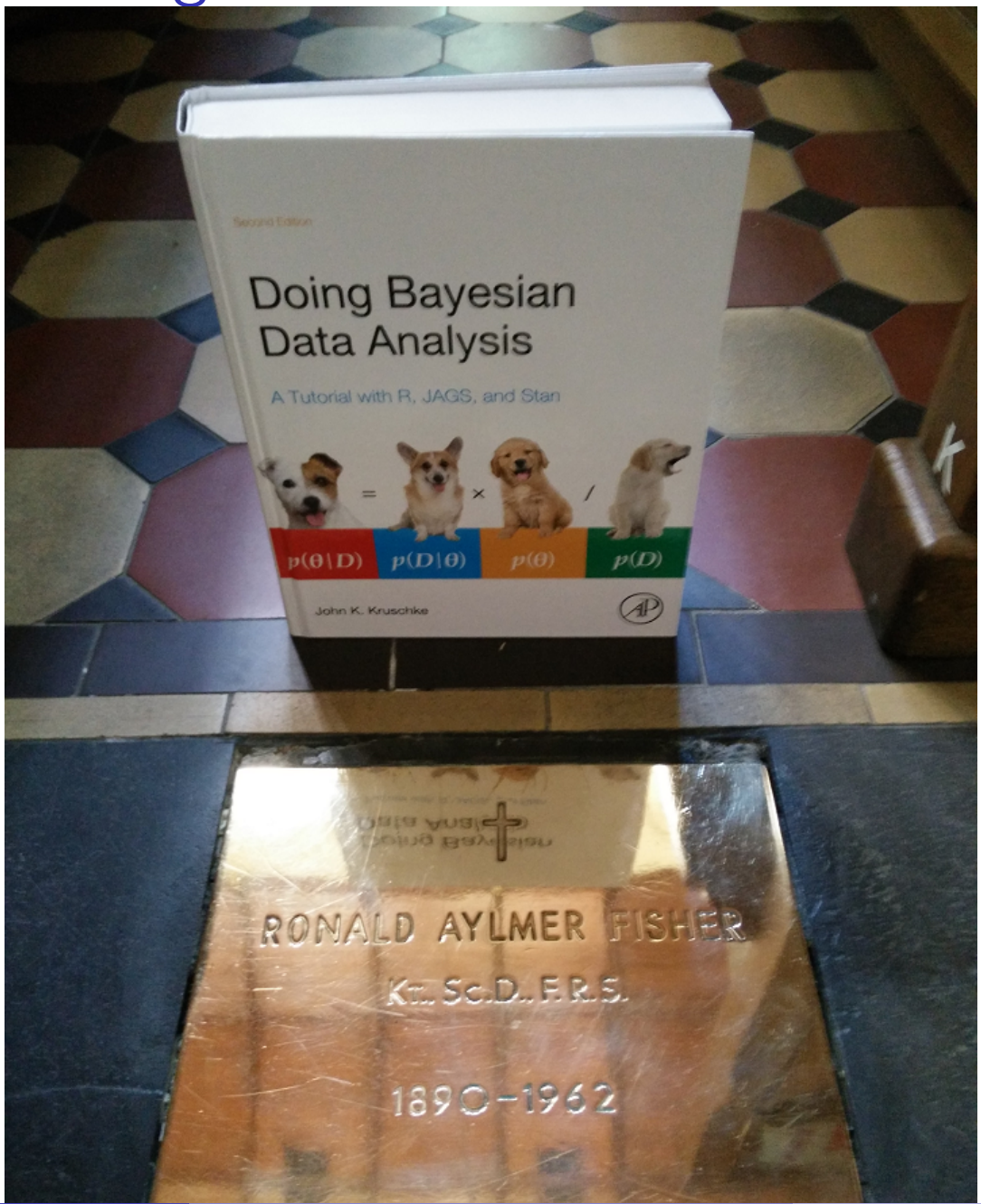
- There is (almost) always a significant result lurking in any given dataset just by chance
  - "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011)
  - "garden of forking paths" (Gelman and Loken 2013)

- We needn't care about significance; we must care about *scientific truth*
- Focus on accurately estimating your effects, making falsifiable predictions and then testing those predictions, and the rest will follow (cf. Wagenmakers et al. 2012; Kruschke 2013; Cumming 2014; Gelman 2015a; Button 2016; Yarkoni and Westfall 2016)
- And if your results are falsified, then we've still learned something

And that's what counts.

The End

palday.bitbucket.org



References

Allefeld, Carsten, Kai Görgen, and John-Dylan Haynes. 2016. "Valid Population Inference for Information-Based Imaging: From the Second-Level T-Test to Prevalence Inference." *NeuroImage* 141: 378–92. doi:10.1016/j.neuroimage.2016.07.040.

Button, Katherine S, John P A Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma S J Robinson, and Marcus R Munafò. 2013. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nat Rev Neurosci*, Apr. doi:10.1038/nrn3475.

Button, Katherine S. 2016. "Statistical Rigor and the Perils of Chance." *Eneuro* 3 (4). eneuro. doi:10.1523/ENEURO.0030-16.2016.

Cumming, Geoff. 2014. "The New Statistics: Why and How." *Psychological Science* 20 (10): 1–23. doi:10.1177/0956797613504966.

Gelman, Andrew. 2013. "P Values and Statistical Practice." *Epidemiology* 24 (1): 69–72. doi:10.1097/EDE.0b013e31827886f7.

———. 2015a. "Statistics and the Crisis of Scientific Replication." *Significance*.

———. 2015b. "The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective." *Journal of Management* 41 (2): 632–43. doi:10.1177/0149206314525208.

Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'P-Hacking' and the Research Hypothesis Was Posited Ahead of Time." `http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf`.

Gelman, Andrew, and Hal Stern. 2006. "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant." *The American Statistician* 60 (4): 328–31. doi:10.1198/000313006X152649.

Jaeger, T. Florian. 2008. "Categorical Data Analysis: Away from ANOVAs (Transformation or Not) and Towards Logit Mixed Models." *Journal of Memory and Language* 59 (4): 434–46. doi:10.1016/j.jml.2007.11.007.

Kruschke, John K. 2013. "Bayesian Estimation Supersedes the T Test." *J Exp Psychol Gen* 142 (2): 573–603. doi:10.1037/a0029146.

Murdoch, Duncan J, Yu-Ling Tsai, and James Adcock. 2008. "*P*-Values Are Random Variables." *The American Statistician* 62 (3): 242–45.

Nieuwenhuis, Sander, Birte U Forstmann, and Eric-Jan Wagenmakers. 2011. "Erroneous Analyses of Interactions in Neuroscience: A Problem of Significance." *Nat Neurosci* 14 (9): 1105–7. doi:10.1038/nn.2886.

Sassenhagen, Jona, and Phillip M. Alday. 2016. "A Common Misapplication of Statistical Inference: Nuisance Control with Null-Hypothesis Significance Tests." *Brain and Language* 162 (November): 42–45. doi:10.1016/j.bandl.2016.08.001.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. doi:10.1177/0956797611417632.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Hans L J van der Maas, and Rogier A Kievit. 2012. "An Agenda for Purely Confirmatory Research." *Perspectives on Psychological Science* 7 (6): 632–38. doi:10.1177/174569161246307.

Wilcox, Rand R. 2010. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*. 2. New York: Springer. doi:10.1007/978-1-4419-5525-8.

Yarkoni, Tal, and Jacob Westfall. 2016. "Choosing Prediction over Explanation in Psychology: Lessons from Machine Learning." *FigShare*. figshare. doi:10.6084/m9.figshare.2441878.v1.