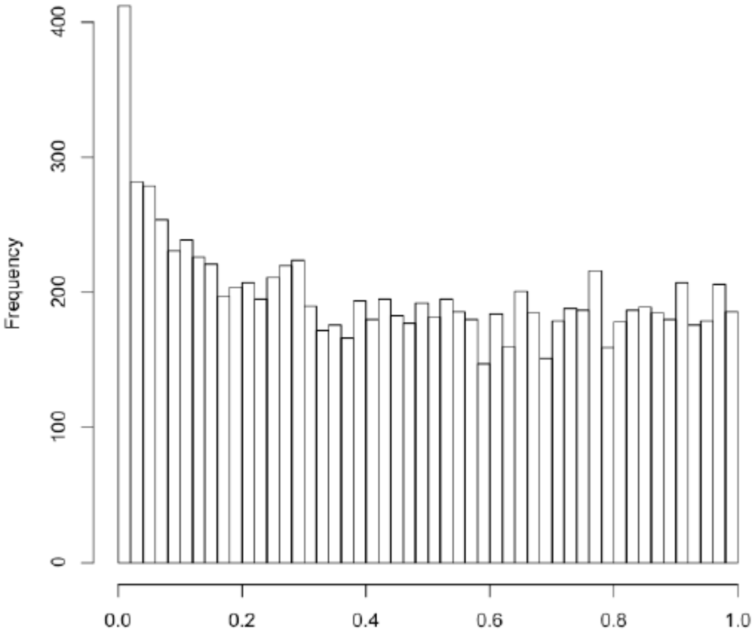
Lecture 11:
P-values revisited

BMI 713
November 28, 2017
Peter J Park

Multiple Testing Correlations

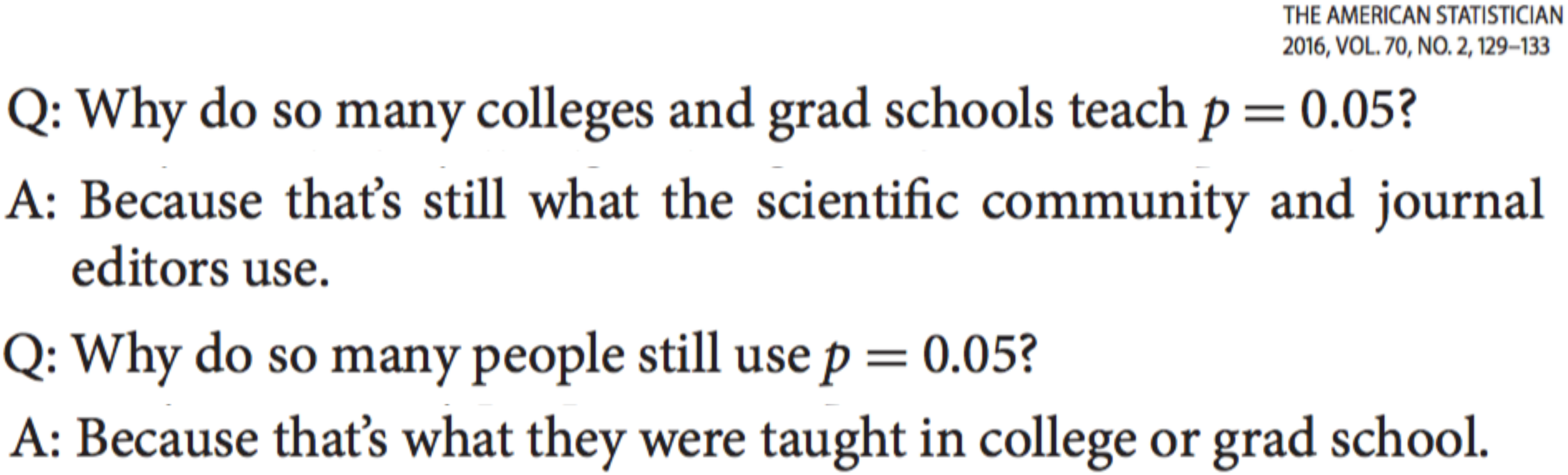
- So what is the procedure in practice?
- Should the significance of my gene depend on that of other genes?

Plot the distribution of p-values



- What is the proper threshold for q-values?

Problems with p-value



THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133

Q: Why do so many colleges and grad schools teach $p = 0.05$?

A: Because that's still what the scientific community and journal editors use.

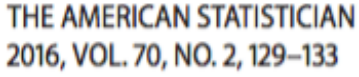
Q: Why do so many people still use $p = 0.05$?

A: Because that's what they were taught in college or grad school.

- "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation." (Siegfried, Science News, 177, 2010)
- One journal has banned its use…

Problems with p-value

- People use the P value to measure the strength of the evidence that the observed value is not a chance occurrence.
- P value is often misinterpreted as the probability that the null hypothesis H_0 is true.
- We assumed that H_0 is true and drew samples from H_0.
- **Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value.**
- A small P-value means that an improbable event has occurred in the context of this assumption.
- Why p=0.05?
- What is the proper null distribution?



THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133

“ASA Statement on Statistical Significance and P-Values”

- **P-values can indicate how incompatible the data are with a specified statistical model.** This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.
- **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.** It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.
- **Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.** Pragmatic considerations often require binary, “yes-no” decisions, but this does not mean that *p*-values alone can ensure that a decision is correct or incorrect.

continued…

- **Proper inference requires full reporting and transparency** Avoid cherry-picking, data dredging, significance chasing, "p-hacking", etc.
- **A p-value, or statistical significance, does not measure the size of an effect or the importance of a result**. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough.
- **By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.** For example, a relatively large p-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data.

So what should one do?

- Supplement or replace p-values with other approaches
- **Use of confidence intervals (CIs) provide information about precision as well as statistical significance**
- Other methods: Bayesian methods, likelihood ratios, false discovery rates, etc.

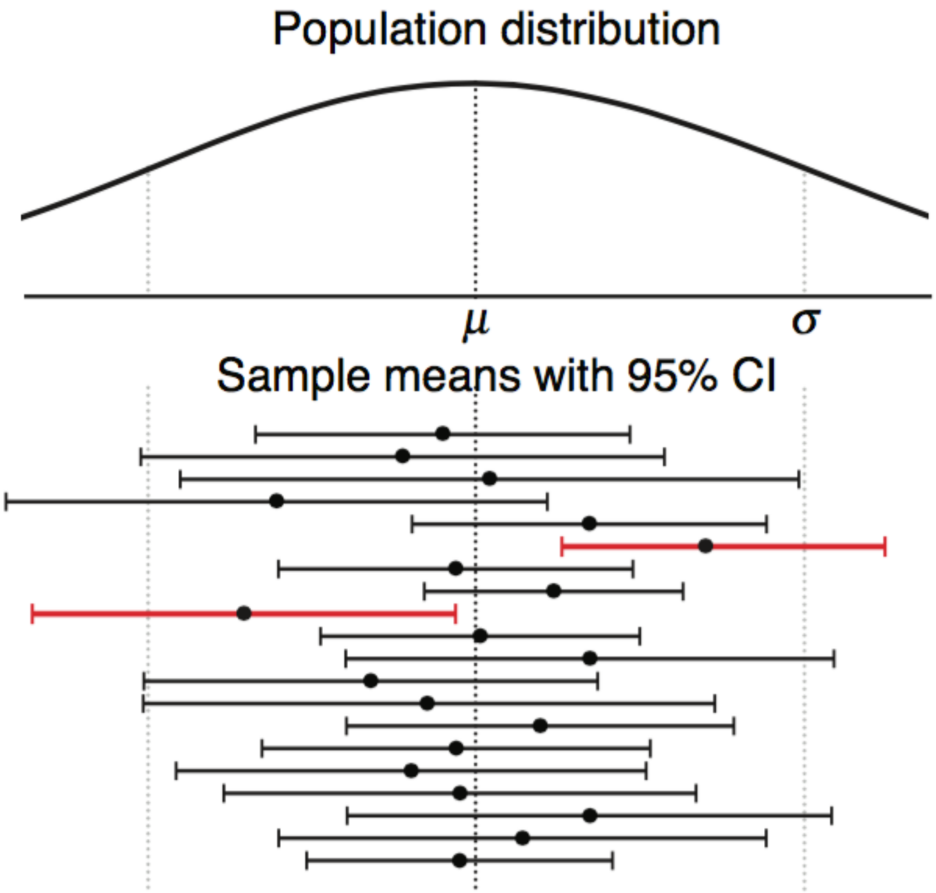
- But do people understand CIs and standard error (SE) bars?

Error bars

- A survey in Nature Methods
  - Error bars found in about two-thirds of the figure panels in which they could be expected (scatter and barplots).
  - s.d. bars: 45%
  - s.e.m. bars: 49%
  - 5%: the error bar type was not specified in the legend.
  - 95% CI was rare
- But CIs are a more intuitive measure of uncertainty and are popular in the medical literature.

Confidence intervals - meaning?

- T/F: The specific 95% confidence interval presented by a study has a 95% chance of containing the mean.



- The 95% CI captures the population mean 95% of the time.

Belia et al, Researchers Misunderstand Confidence Intervals and Standard Error Bars, Psychological Methods, 2005

Interpreting error bars: CIs

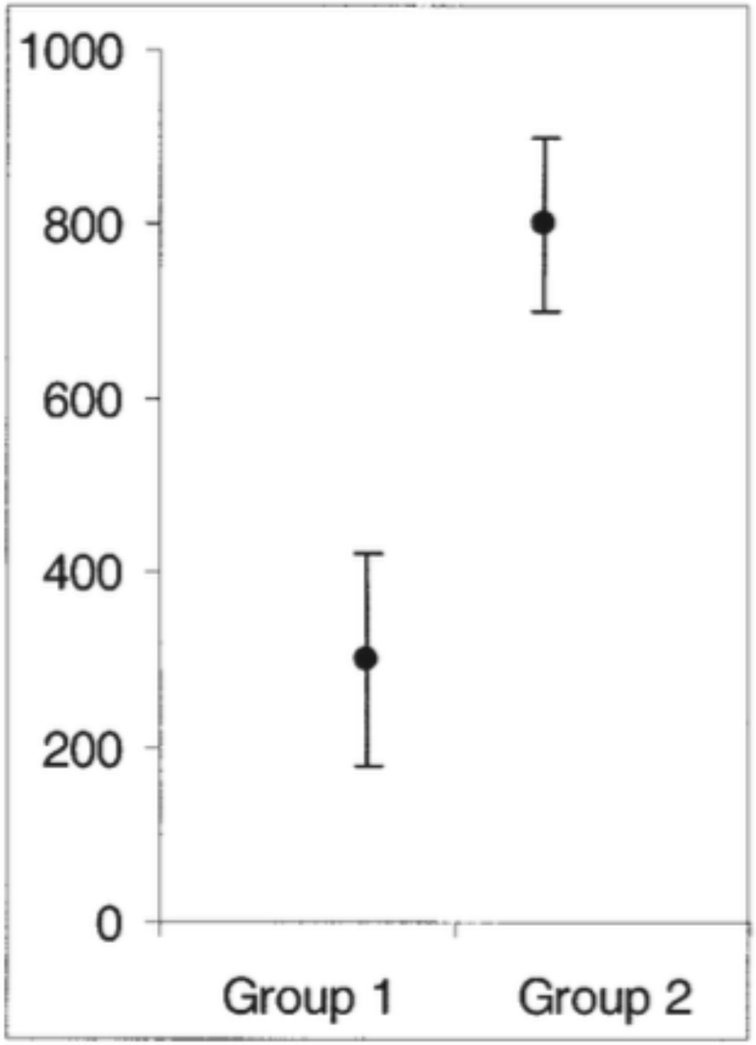
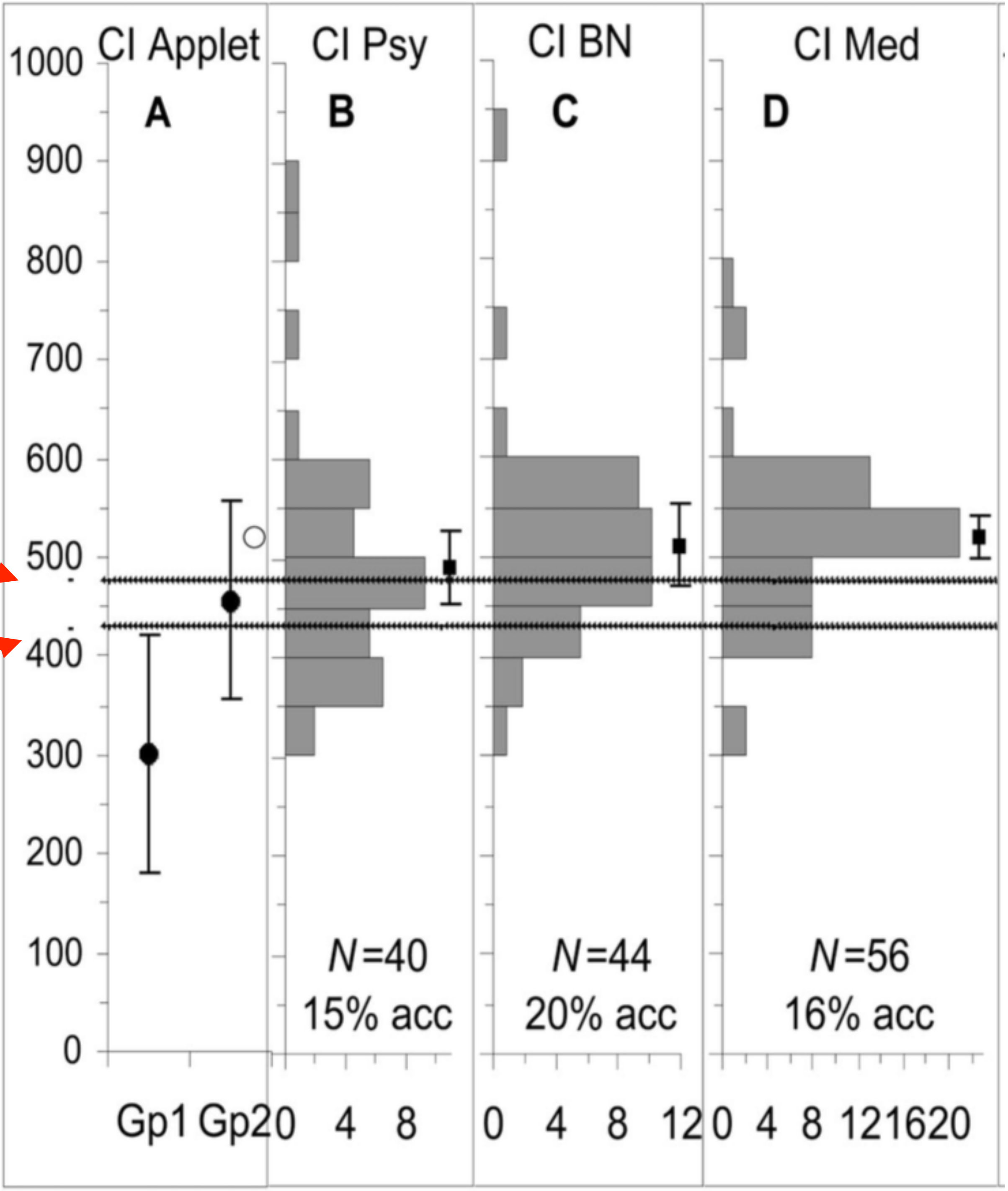


Figure 1. Mean reaction time (ms) and 95% Confidence Intervals for Group 1 (n=36) and Group 2 (n=34).

TASK: move the second error bar until the two means are just significantly different (t-test, two-tailed, p<0.05)

Psy - psychology; BN - behavioral neuroscience; Med - medicine

Emailed ~4000 authors from 978 articles (1999-2001) in 33 leading journals, 15% responded



0.025

0.1

Under some conditions:

~1/4 overlap of CIs -> p-value of 0.05

Overlapping intervals?

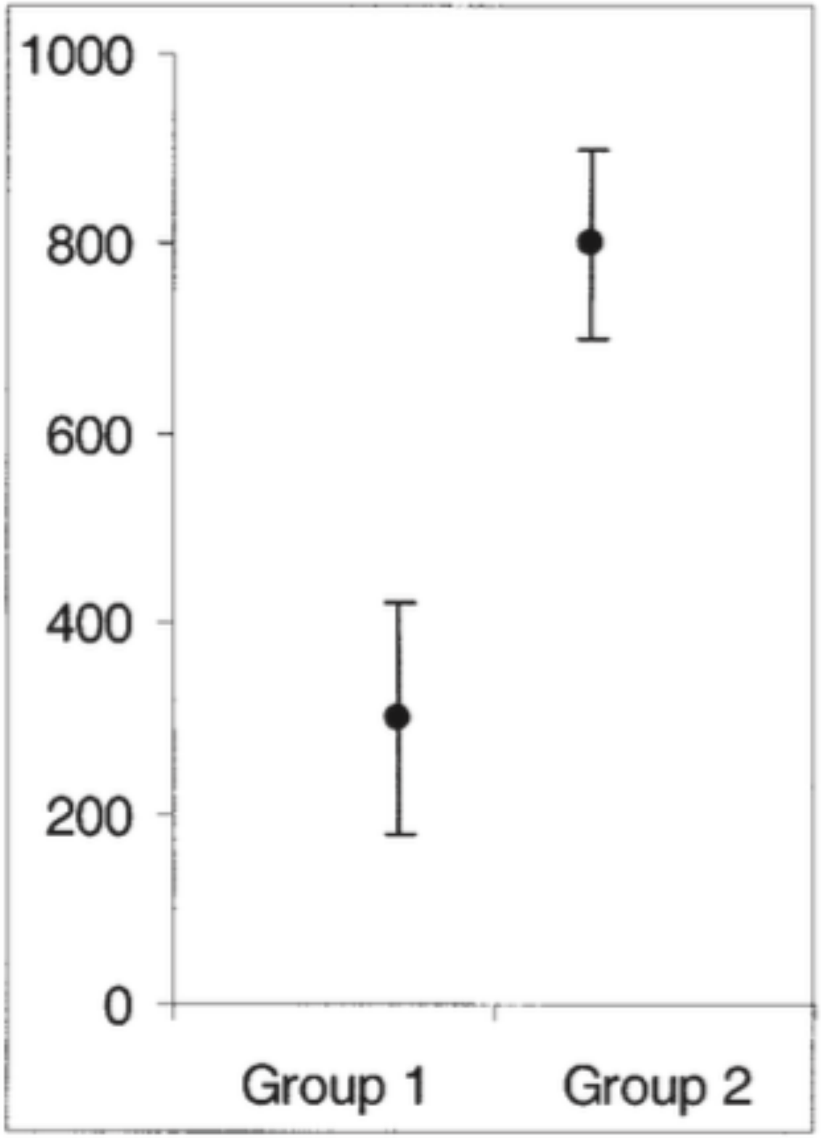
- A widely believed rule: overlap of the two CIs implies there is no significant difference.
- In fact, nonoverlap of the two CIs does imply a significant difference, but with p-value much less than .05

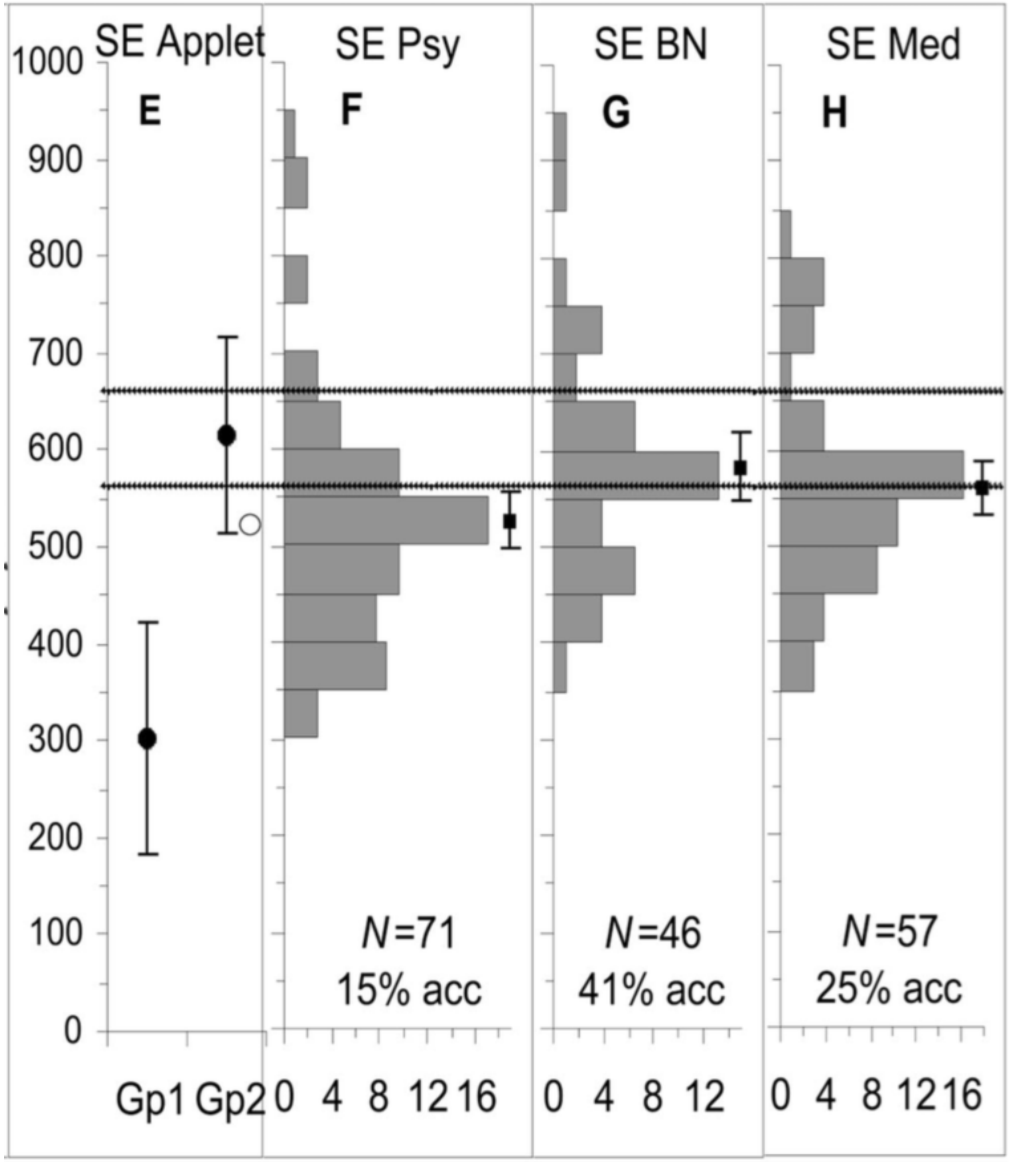
- There was an anchoring effect!

Interpreting error bars: SEs

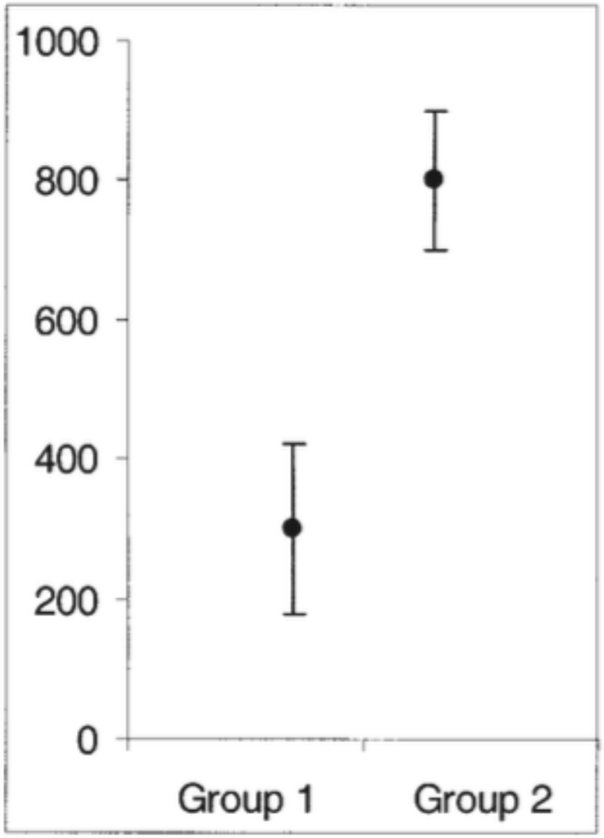
Under some conditions:

Gap is ~average of SEs -> p-value of 0.05

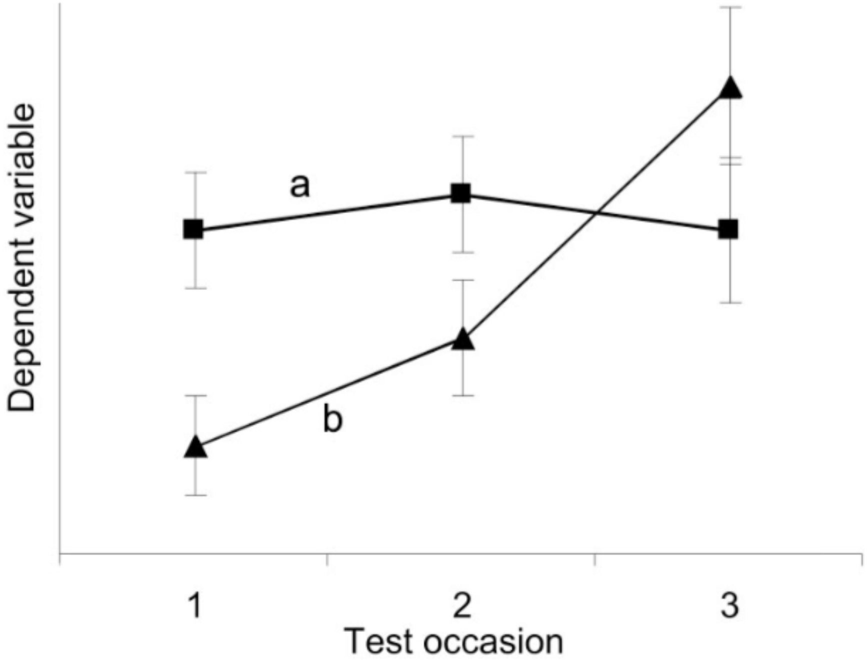




Repeated measurements



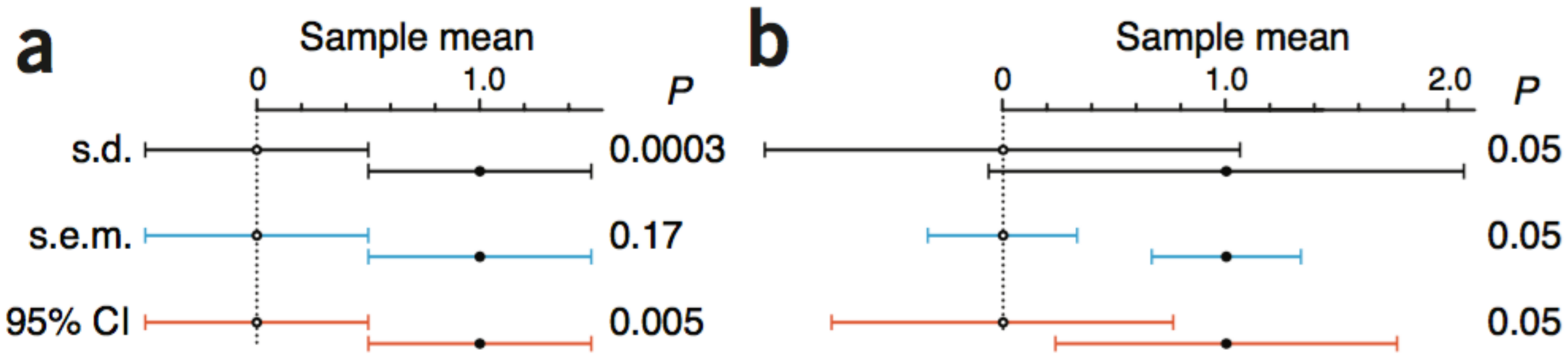
For another testers, groups were marked with “pre-test” and “post-test” and it was noted that the data were for a single group

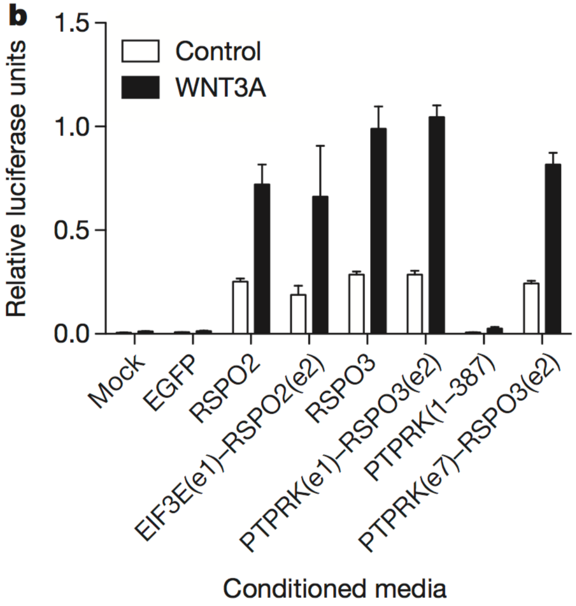


Observations

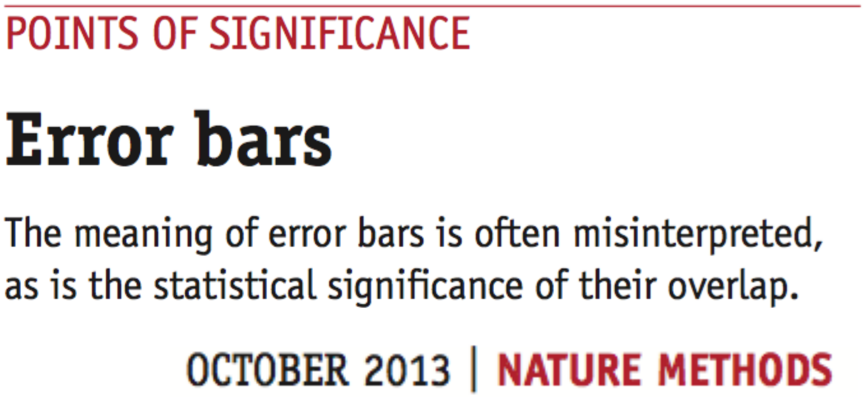
- Only 22% set the means so the p-value was between .025 and .10.
- Respondents overall did not adequately distinguish CIs and SE bars!
- Many respondents (31.5%) used the incorrect rule that error bars, whether a 95% CI or SE bars, just touch when means are just statistically significantly different (p=.05).
- A large majority overlooked the clear statements that the means they saw were from a repeated measures or paired design.

When the different error bars touch…





$(n = 10)$



POINTS OF SIGNIFICANCE

Error bars

The meaning of error bars is often misinterpreted, as is the statistical significance of their overlap.

OCTOBER 2013 | NATURE METHODS

Seshagiri et al, Nature

“Error bars represent mean +/-s.d. from three replicate experiments”

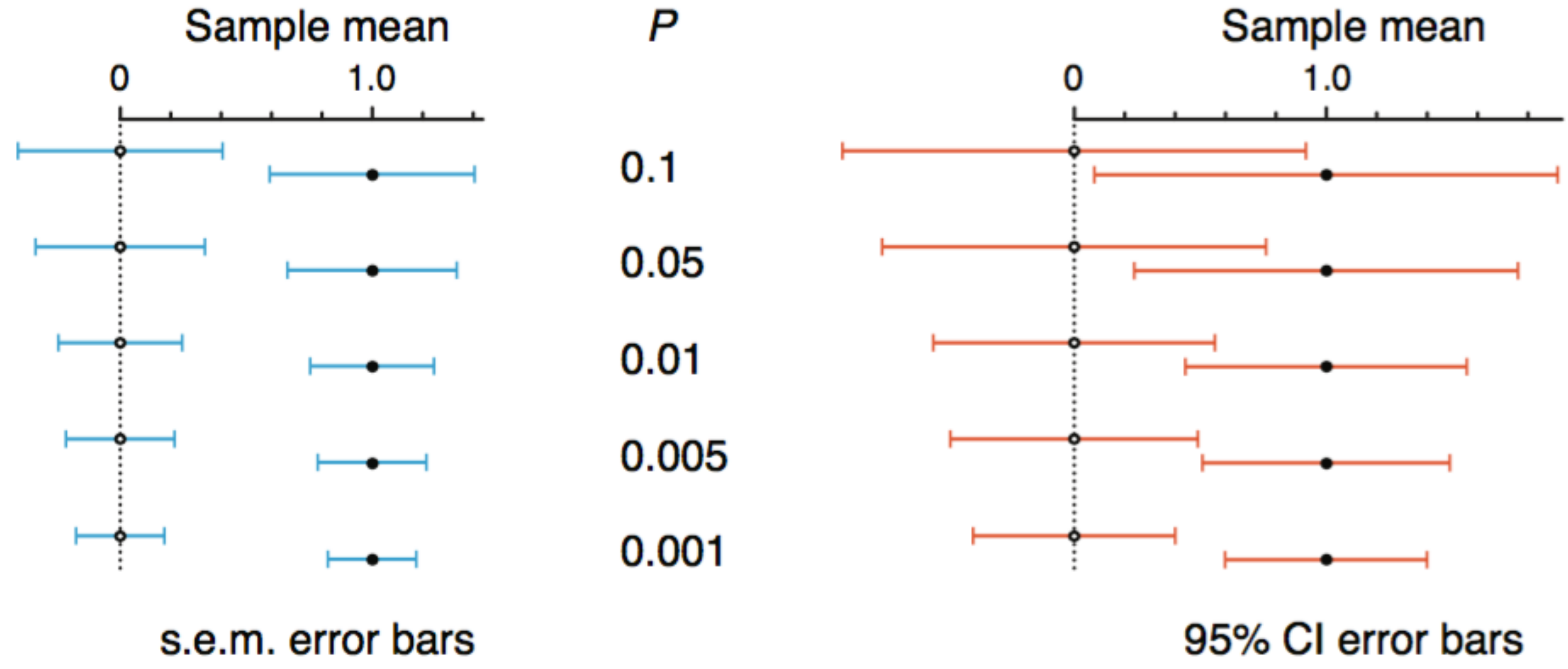
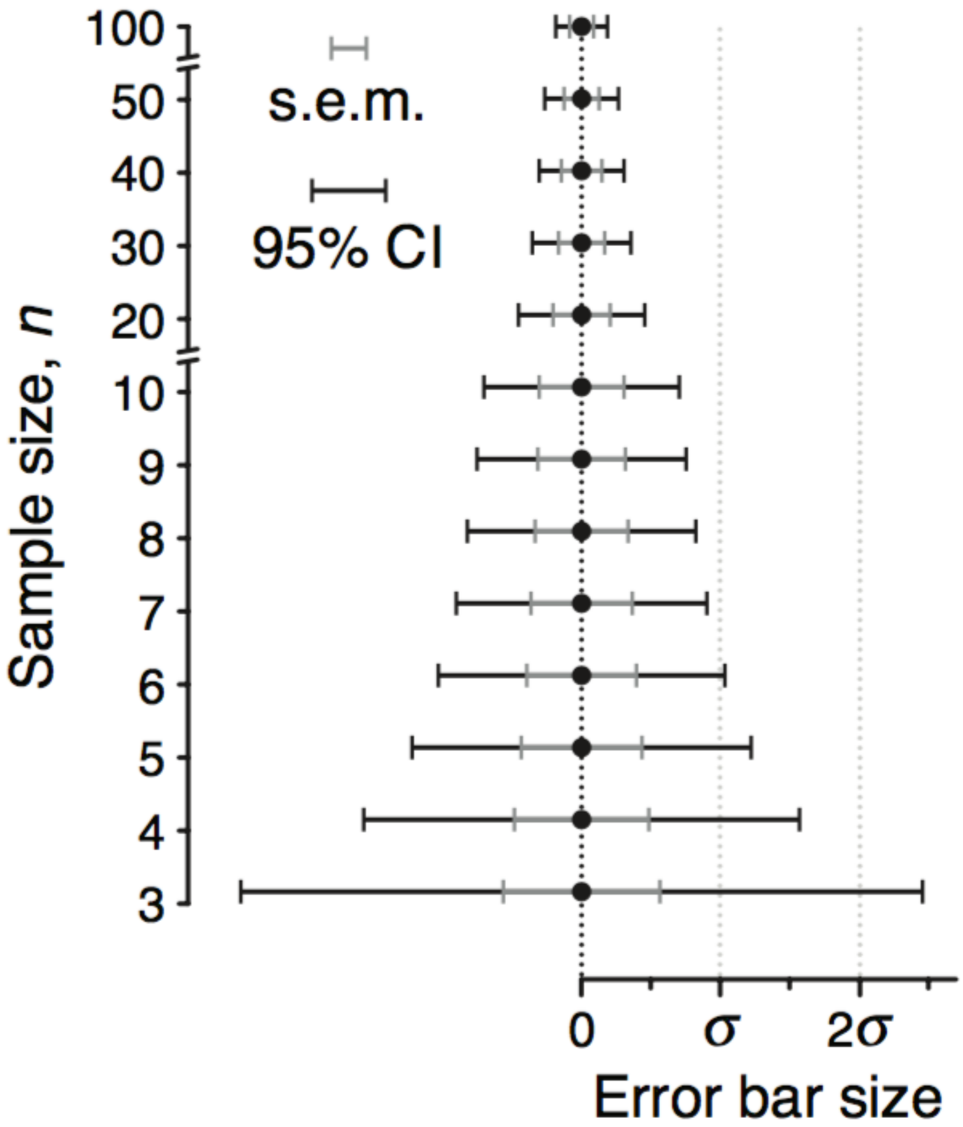
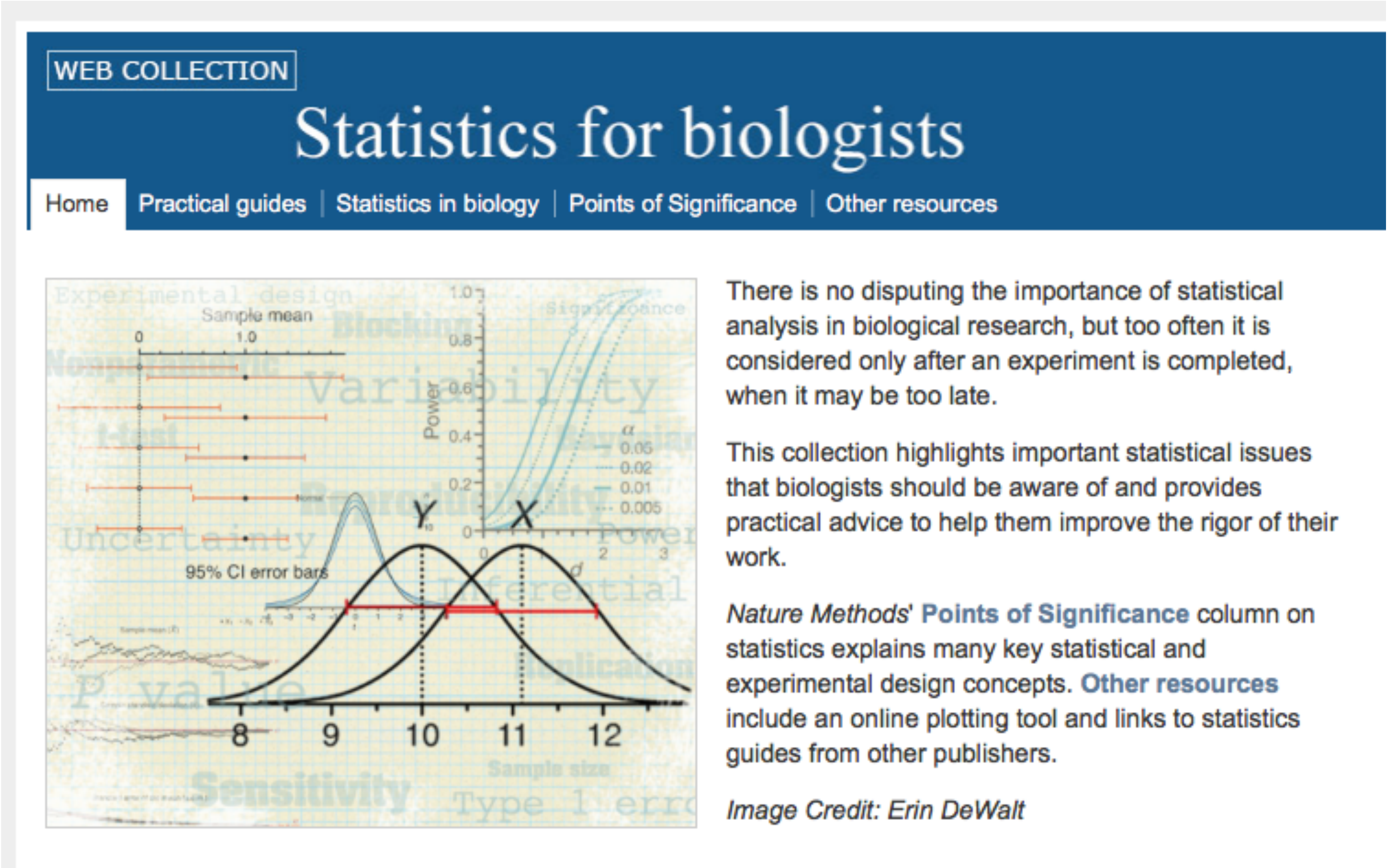


Figure 3 | Size and position of s.e.m. and 95% CI error bars for common *P* values. Examples are based on sample means of 0 and 1 ($n = 10$).

Impact of sample size on s.e.m.



The two are related by the t-statistic.



Statistics for biologists

There is no disputing the importance of statistical analysis in biological research, but too often it is considered only after an experiment is completed, when it may be too late.

This collection highlights important statistical issues that biologists should be aware of and provides practical advice to help them improve the rigor of their work.

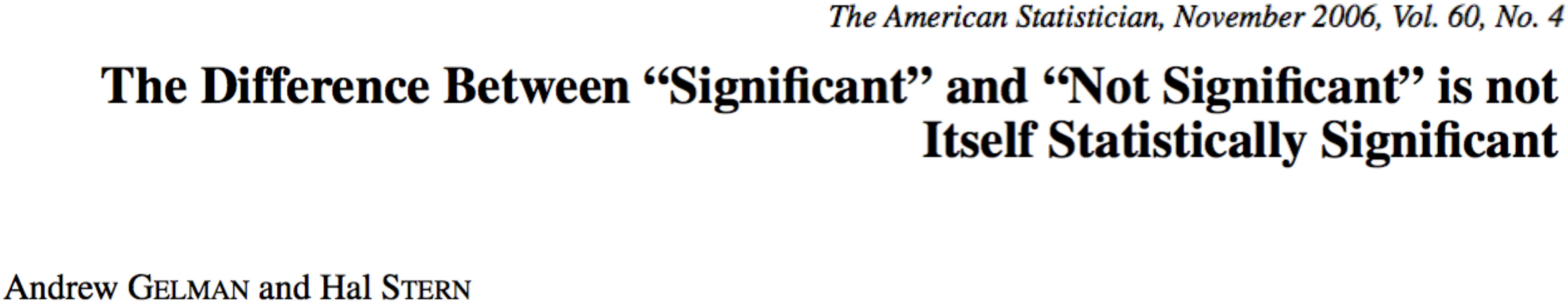
Nature Methods' **Points of Significance** column on statistics explains many key statistical and experimental design concepts. **Other resources** include an online plotting tool and links to statistics guides from other publishers.

Image Credit: Erin DeWalt

A MAJOR COMMON MISTAKE

- You want to show that training effect on neuronal activity in mutant mice is different from its effect in control mice

- “The percentage of neurons showing cue-related activity increased with training in the mutant mice ($P < 0.05$), but not in the control mice ($P > 0.05$)”



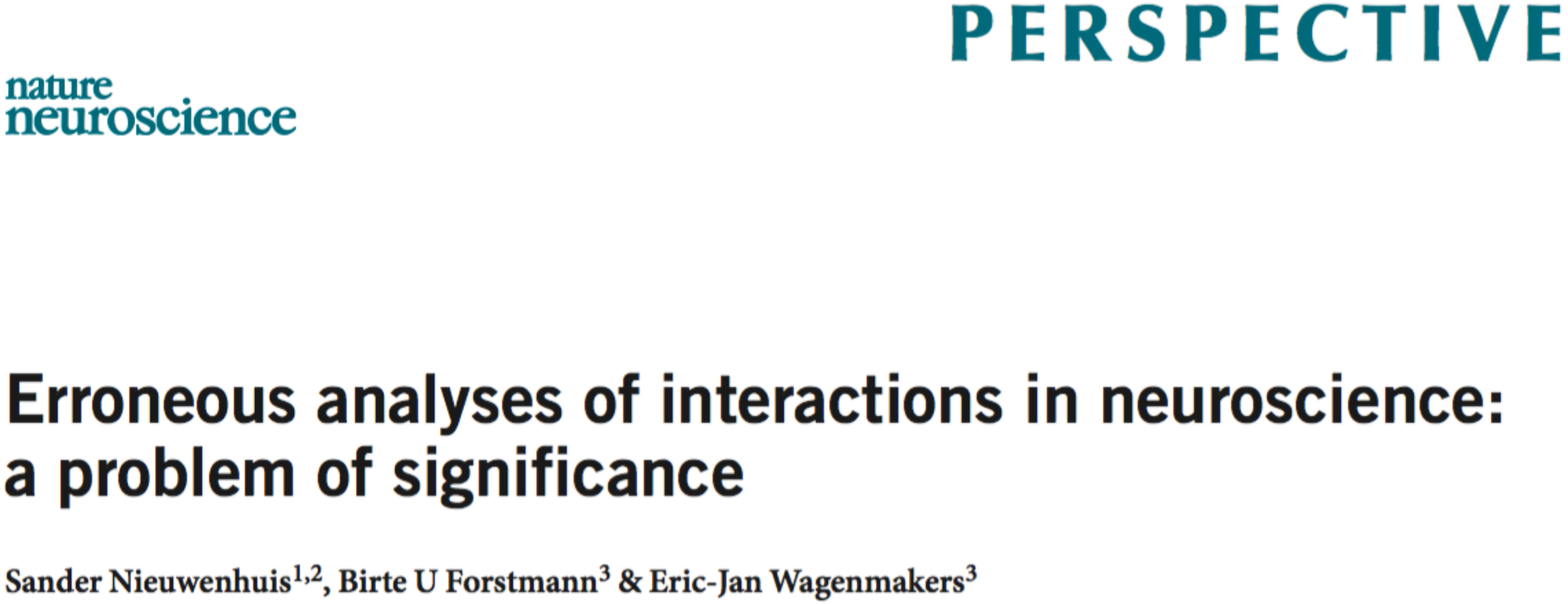
The American Statistician, November 2006, Vol. 60, No. 4

The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant

Andrew GELMAN and Hal STERN

Two problems

- Problem with binarization: in an extreme case, one might be significant with P = 0.049 while the other barely misses significance with P = 0.051.
- Consider the effect of drugs on blood pressues in these three scenarios with the following effect estimates and s.e.:
    1. 25 +/- 10 (significant at 1%)
    2. 10 +/- 10 (not significant)
    3. 2.5 +/- 1.0
- Is the difference between #1 and #2 significant? (15 with s.e.=14)
- Is the difference between #1 and #3 significant?
- Did the third study replicate the first study? (effect size or significance?)



PERSPECTIVE

nature neuroscience

Erroneous analyses of interactions in neuroscience: a problem of significance

Sander Nieuwenhuis[1,2], Birte U Forstmann[3] & Eric-Jan Wagenmakers[3]

- Question: Can you compare two results by comparing their degree of statistical significance?

From the abstract

In theory, a comparison of two experimental effects requires a statistical test on their difference. In practice, this comparison is often based on an incorrect procedure involving two separate tests in which researchers conclude that effects differ when one effect is significant ($P < 0.05$) but the other is not ($P > 0.05$).

- Reviewed 513 articles: all of the behavioral, systems and cognitive neuroscience studies published in five prestigious journals
- 157 (31%) articles describe at least one situation in which they might be tempted to make the error.
- 78 used the correct procedure and 79 used the incorrect procedure

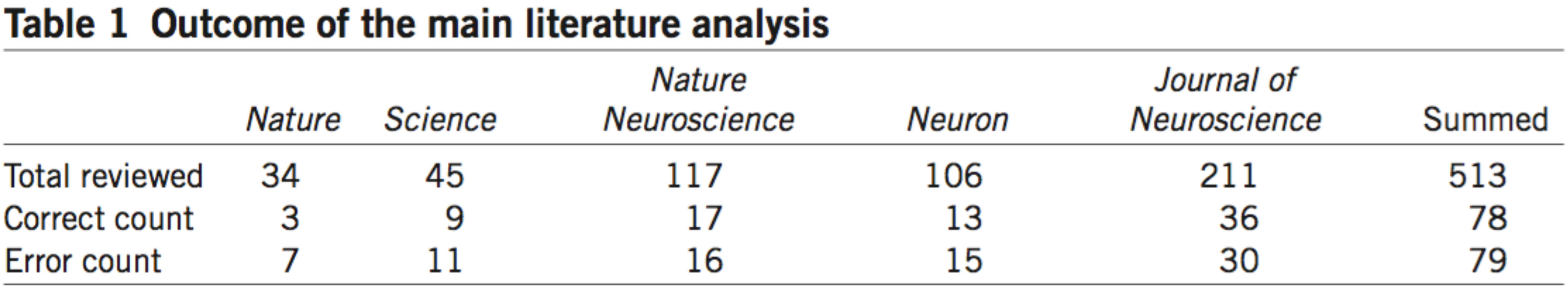


Table 1 Outcome of the main literature analysis

	Nature	*Science*	*Nature Neuroscience*	*Neuron*	*Journal of Neuroscience*	Summed
Total reviewed	34	45	117	106	211	513
Correct count	3	9	17	13	36	78
Error count	7	11	16	15	30	79

Are all these articles wrong about their main conclusions?

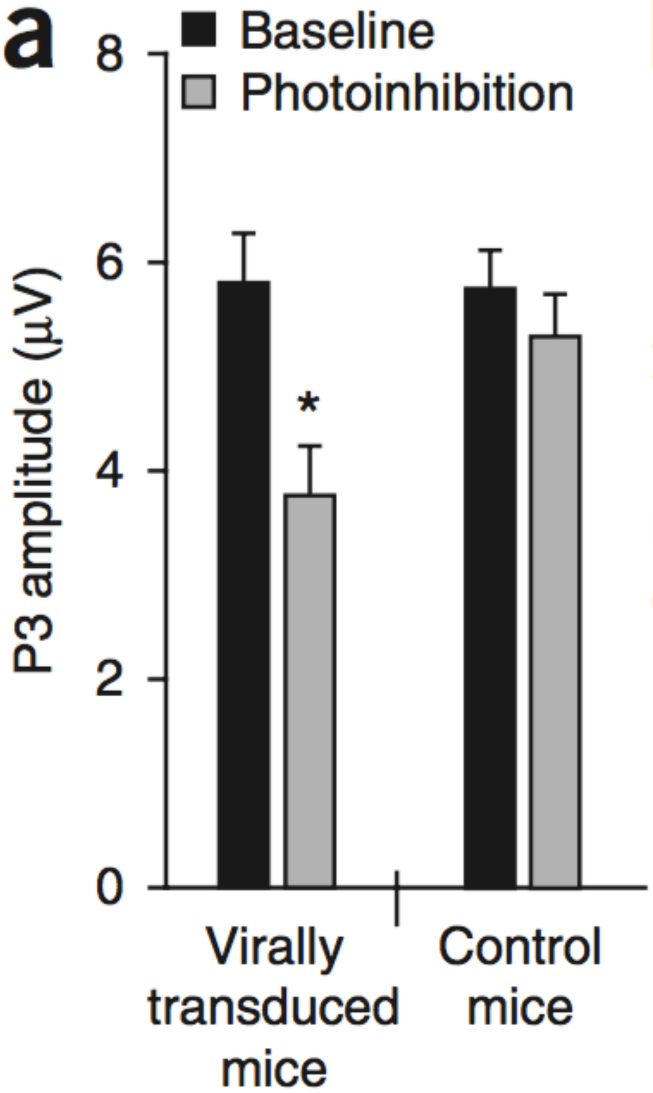
- For a given paper, the main conclusions may not depend on the erroneous analysis
- In ~1/3, “we were convinced that the critical, but missing, interaction effect would have been statistically significant (consistent with the researchers’ claim), either because there was an enormous difference between the two effect sizes or because the reported methodological information allowed us to determine the approximate significance level.”
- In ~2/3, not enough information

Cellular/Molecular Neuroscience?

- Additional 120 cellular and molecular neuroscience articles published in *Nature Neuroscience* in 2009 and 2010 (the first five Articles in each issue)
- Not a single study that used the correct statistical procedure to compare effect sizes.
- At least 25 studies that used the erroneous procedure and explicitly or implicitly compared significance levels.
- In general, data were analyzed with t-tests (possibly corrected for multiple comparisons or unequal variances) and occasionally with one-way ANOVAs, even when the experimental design was multifactorial and required a more sophisticated statistical analysis.

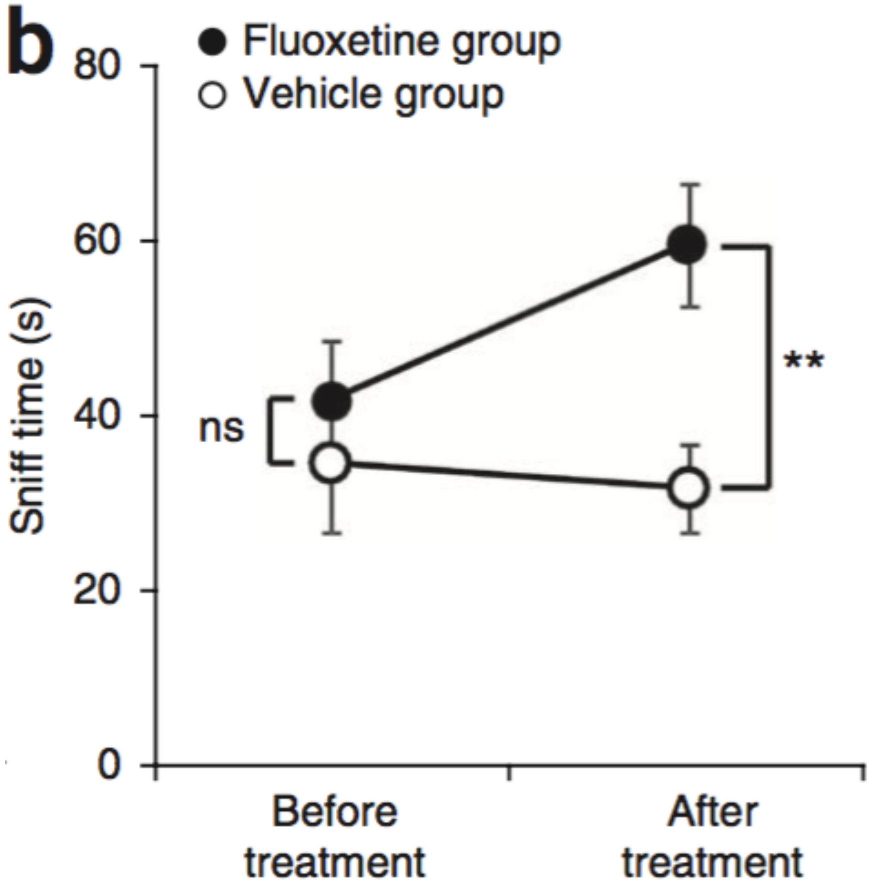
What type of errors?

- Most of the errors when comparing effect sizes in an experimental group/ condition and a control group/ condition (for example, sham-TMS, vehicle infusion, placebo pill, wild-type mice)
- “Optogenetic photoinhibition of the locus coeruleus decreased the amplitude of the target-evoked P3 potential in virally transduced animals (P = 0.012), but not in control animals (P = 0.3)”. Authors’ own data; interaction term not significant.



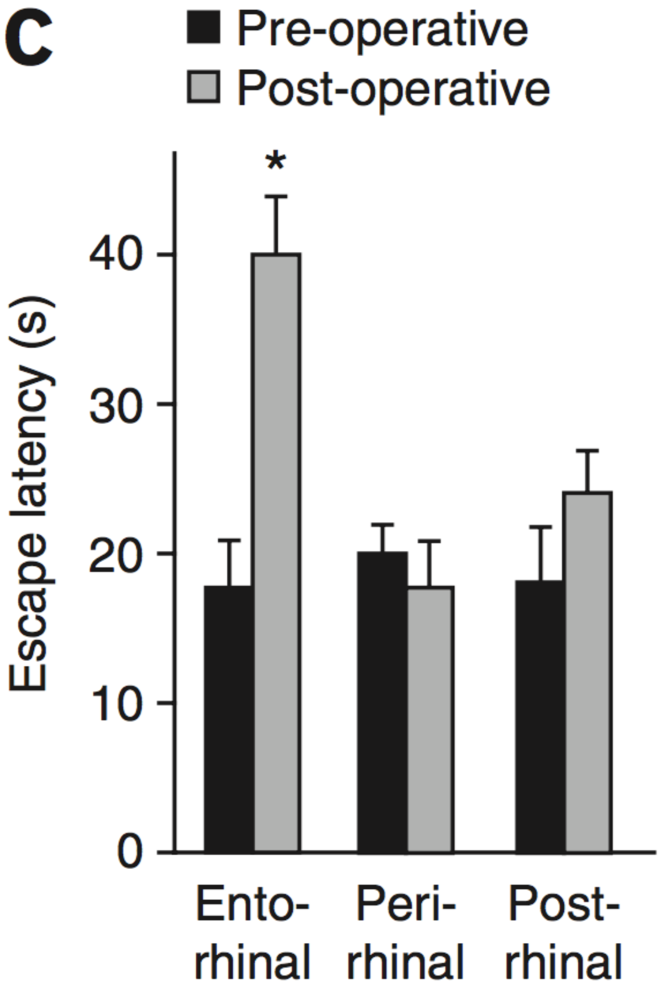
Pre vs post-test

- A special case: comparing effect sizes during a pre-test (control condition) and a post-test (experimental condition)
- Example: “Acute fluoxetine treatment increased social approach behavior (as indexed by sniff time) in our mouse model of depression (P < 0.01)”



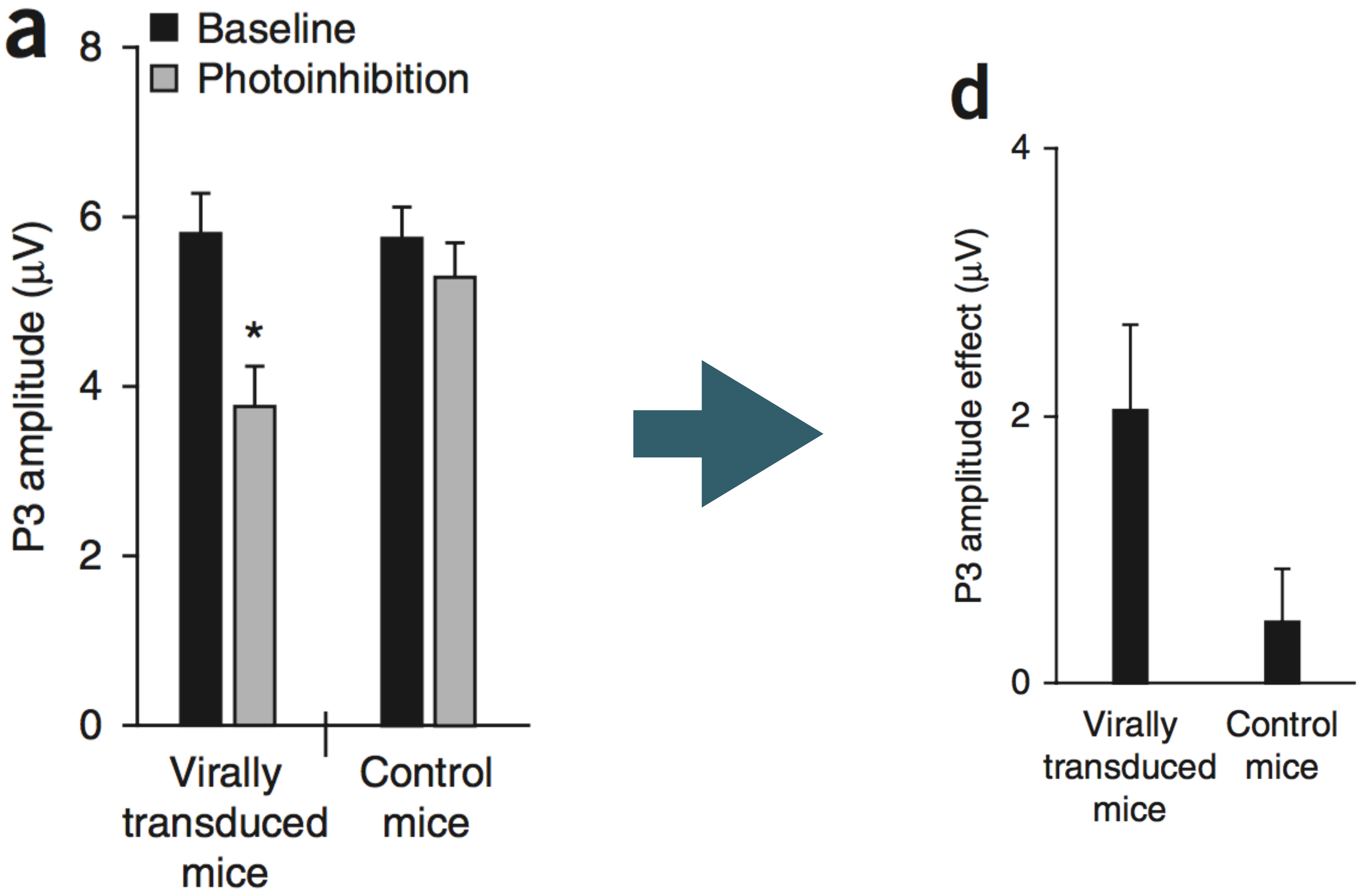
Comparing multiple choices

- Comparing several brain areas and claiming that a particular effect (property) is specific for one of these brain areas.
- Example: “Escape latency in the Morris water maze was affected by lesions of the entorhinal cortex (P < 0.05), but was spared by lesions of the perirhinal and postrhinal cortices (both P values > 0.1), pointing to a specific role for the enthorinal cortex in spatial memory.”



What should you have done?

- Make direct comparisons



Same problem with correlations

- “Hippocampal firing synchrony correlated with memory performance in the placebo condition (r = 0.43, P = 0.01), but not in the drug condition (r = 0.19, P = 0.21)”.
- Again, one should directly compare the two correlations.

What should you have done?

- You need to test for interactions (ANOVA)

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim N\left(0, \sigma^2\right)$$