The Hypothesis Testing Paradox
or
Why Effect Sizes are Important for Evaluating Evidence

Professor Emerita Jessica Utts
Department of Statistics
University of California, Irvine
USCOTS 2019

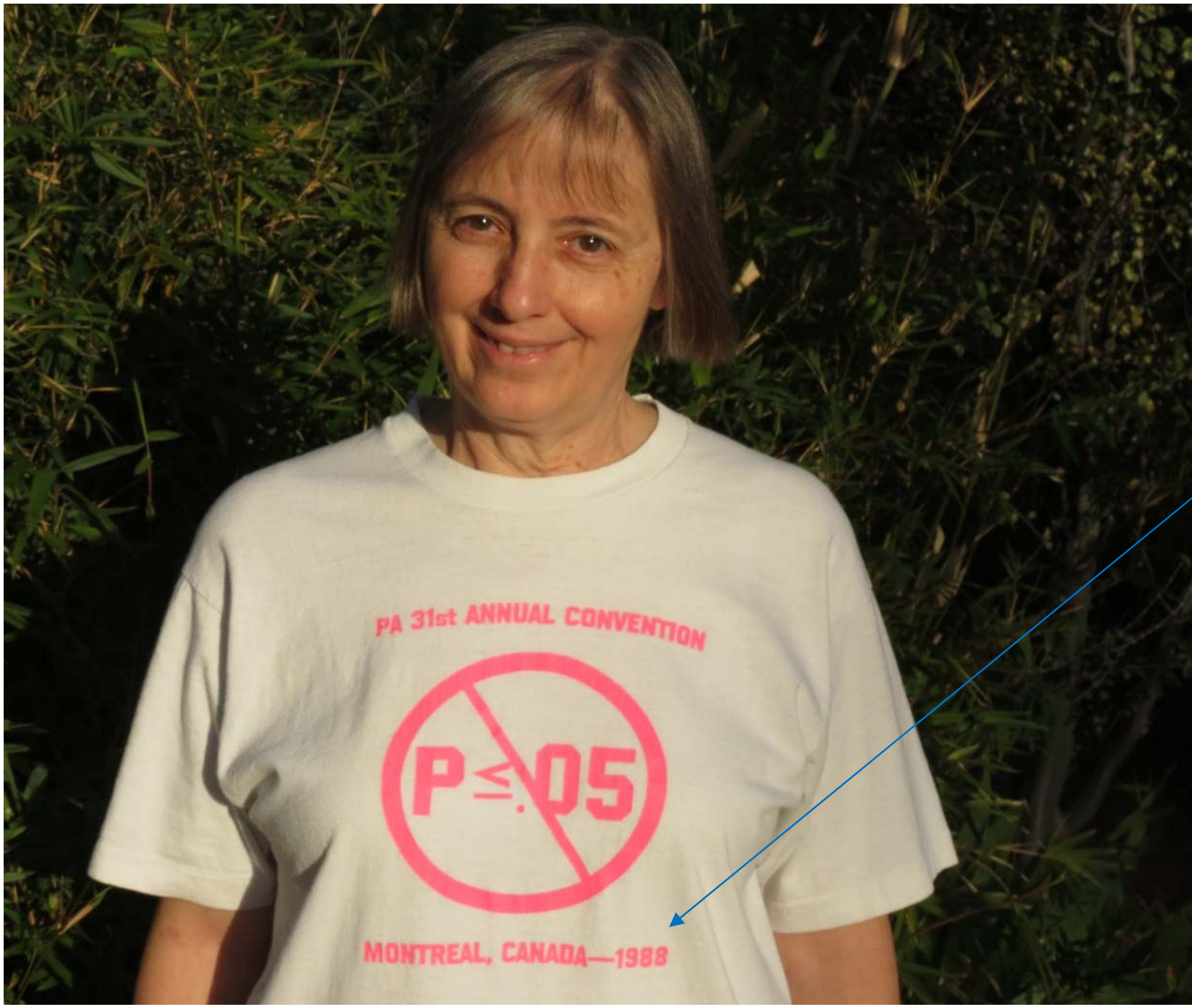
Replicating Research Findings

New NAS report, 2019 (Preprint)

Reproducibility and Replicability in Science:

- "For this report: *Replicability* is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data. (p. 36)"
- "One type of scientific research tool, statistical inference, has an outsized role in replicability discussions due to the frequent misuse of statistics and the use of a *p*-value threshold for determining "statistical significance." (Summary, bullet #7)"

I’ve argued against “statistical significance = successful replication” for a long time!



Notice the date:

1988!

Effect Size Examples

- Test for one population mean:
  - Effect size measures how far true parameter value is from null value, usually in # of standard deviations
- Comparing two population means:
  - Effect size measures difference in means, usually in # of standard deviations for one group
- Example: Average heights for males and females differ by about 5 inches, which is about twice the standard deviation for each sex. So the effect size is about 5/2.5 = 2 (a very large effect)

Example: Are female college students taller than their mothers?

- n = 93 pairs (daughter – mother height)
  - mean difference = 1.3 inches
  - standard deviation = 2.6 inches
- Effect size is 1.3/2.6 = 0.5 (moderate effect)
- Test statistic is $t = \sqrt{93} \times 0.5 = 4.8$, p-value ≈ 0
- Relationship between t and e.s.

$$t = \sqrt{n}\boxed{\left(\frac{\bar{x}-\mu_0}{s}\right)} \qquad e.s. = \boxed{\frac{\bar{x}-\mu_0}{s}} \qquad t = \sqrt{n} \times e.s.$$

Hypothesis testing paradox:

- A researcher conducts a test with *n* = 100 and gets these results:
  - $t = \sqrt{100}\left(\frac{\bar{x}-\mu_0}{s}\right) = 2.50$
  - *p*-value = 0.014, reject null hypothesis
- Just to be sure, the researcher decides to repeat the experiment with *n* = 25

Hypothesis testing paradox:

- Uh-oh, the results show:
  - $t = \sqrt{25}\,\left(\frac{\bar{x}-\mu_0}{s}\right) = 1.25$
  - *p*-value = 0.22, cannot reject null!
  - The effect has disappeared!
- To salvage, researcher decides to combine data:
  - $n = 125$
  - Finds $t = \sqrt{125}\,\left(\frac{\bar{x}-\mu_0}{s}\right) = 2.795$, *p*-value = 0.006!
  - The effect is stronger than the first time!

Hypothesis testing paradox:

- Paradox: The 2nd study *alone* did <u>not</u> "replicate" the finding, but when *combined* with 1st study, the effect seems <u>even stronger</u> than 1st study!
- Defining "replication" as getting statistical significance each time, or on the basis of *p*-values, makes no sense! Yet, it's very common practice in many disciplines.

What's going on?

Study	***n***	**Effect size**	$t = \sqrt{n} \times e.s.$	***P*-value**
1	100	**0.25**	2.50	0.014
2	25	**0.25**	1.25	0.22
Combined	125	**0.25**	2.795	0.006

- In all 3 cases the effect size is the *same*, 0.25.
- But the test statistic and *p*-value change based on the sample size, with $t = \sqrt{n} \times$ (effect size).

Why Effect Sizes are Important

- Unlike *p*-values, they don't depend on sample size (but accuracy of estimating them does).
- They are a measure of the true effect or difference in the population = practical importance!
- Replication should be defined as getting approximately the same effect size, *not* as getting approximately the same *p*-value!