JSM 2010, Session #119

The Undetectable Difference: An Experimental Look at the "Problem" of p-Values

by

William M. Goodman, Ph.D.
University of Ontario Institute of Technology

The Problematic Inequality

P([The sample data have the obtained distribution] | [H_0 is true])

$\neq$

P([H_0 is true] | [The sample data have the obtained distribution])

The Problematic Inequality

P([The sample data have the obtained distribution] | [H_0 is true])

Issue 1: "Thickness" of H_0 ?

$\neq$

P([H_0 is true] | [The sample data have the obtained distribution])

The Problematic Inequality

P([The sample data have the obtained distribution] | [H_0 is true])

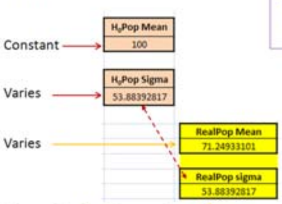
Issue 1: "Thickness" of H_0 ?

$\neq$

Issue 2: Mismatched structures?

P([H_0 is true] | [The sample data have the obtained distribution])

Parameters:



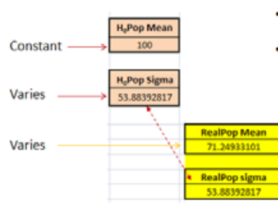
General description of the experiment.

Assumptions for this experiment:
- Real pop'n sigma = H_0Pop Sigma
- Distribution of real population is normal

***Between each experimental pass*:**
- **Randomly vary the two parameters shown above**

(The simulated experimenter will not know how or if these have varied from H_0.)

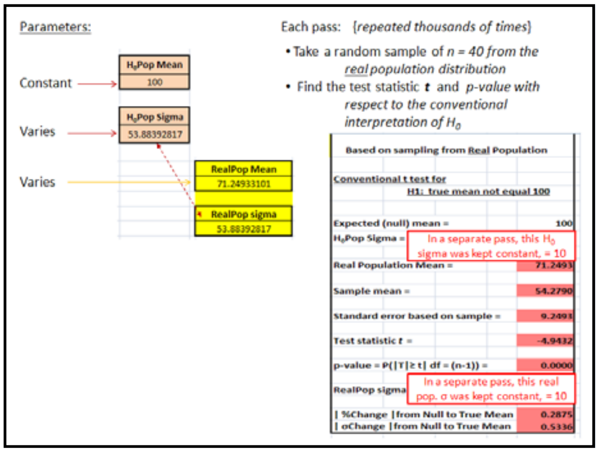
Parameters:



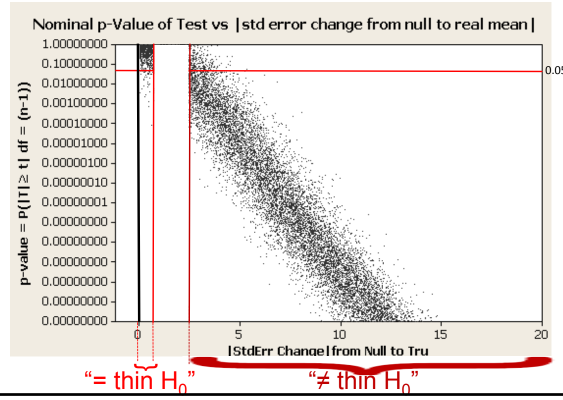
Each pass: {*repeated thousands of times*}
- Take a random sample of *n = 40 from the real population distribution*
- Find the test statistic *t* and *p-value with respect to the conventional interpretation of H_0*

To normalize the distance measures between the H_0 mean and the true mean, I adapted the use of "standard errors":
- Units based on **(true pop'n *σ*) / (√*n*)**
- i.e. analogous to the measure used for constructing a confidence interval (*σ known*)

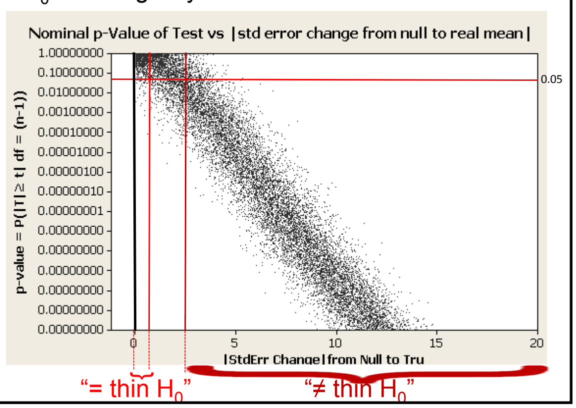
Based on sampling from Real Population	
Conventional t test for H1: true mean not equal 100	
Expected (null) mean =	100
H_0Pop Sigma =	53.8839
Real Population Mean =	71.2493
Sample mean =	54.2790
Standard error based on sample =	9.2493
Test statistic *t* =	-4.9432
p-value = P(|T|≥ t| df = (n-1)) =	0.0000
RealPop sigma =	53.8839
|| σChange |from Null to True Mean	0.5336
| Standard Error change | from Null to True Mean	3.375



If H_0 were digitally thin……..

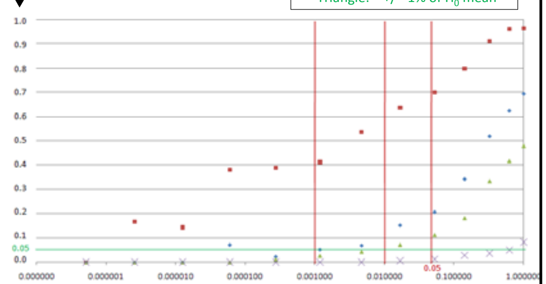


If H_0 were digitally thin……..*which it's not*



Pr(the true population mean ≈ H_0 mean | (the p-value returned from the data is in the indicated range))

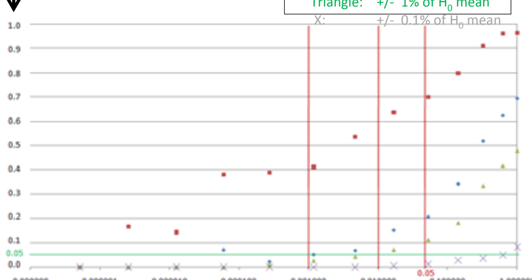
"Approximately equal to H_0 mean":
Box: +/- 2 standard errors
Diamond: +/- 1 standard error
Triangle: +/- 1% of H_0 mean



Approximate p-Value based on a t test for μ = (null mean).
n = 40. H_0 mean = 100. True mean varies. True σ held at 10.

Pr(the true population mean ≈ H_0 mean | (the p-value returned from the data is in the indicated range))

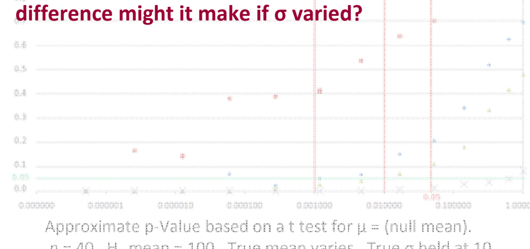
"Approximately equal to H_0 mean":
Box: +/- 2 standard errors
Diamond: +/- 1 standard error
Triangle: +/- 1% of H_0 mean
X: +/- 0.1% of H_0 mean



Approximate p-Value based on a t test for μ = (null mean).
n = 40. H_0 mean = 100. True mean varies. True σ held at 10.

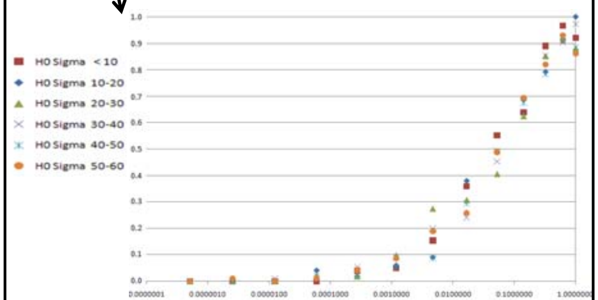
At least two questions arise:

1. **The graph suggests there might be a "thickness" of H_0 at which p-value = 0.05 *and* P(H_0| p-value) = 0.05. What is that point, and is it generalizable???**
2. **In this simulation σ was held constant. What difference might it make if σ varied?**



Pr(the true population mean ≈ H_0 mean | (the p-value returned from the data is in the indicated range))

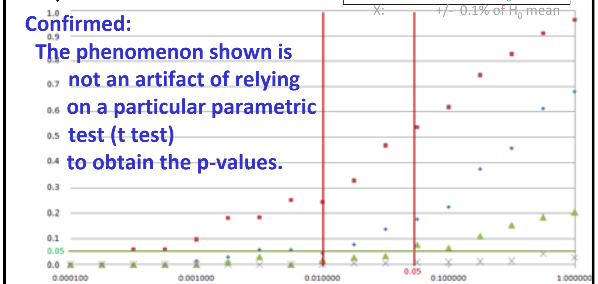
"Approximately equal to H_0 mean":
(All cases: +/- 2 standard errors)
20,000 re-samples



Approximate p-Value based on a t test for μ = (null mean).
n = 40. H_0 mean = 100. True mean varies. True σ varies.

Pr(the true population mean ≈ H_0 mean | (the p-value returned from the data is in the indicated range))

"Approximately equal to H_0 mean":
Box: +/- 2 standard errors
Diamond: +/- 1 standard error
Triangle: +/- 1% of H_0 mean
X: +/- 0.1% of H_0 mean



Approximate, resampling-based p-Value for μ = (null mean).
n = 40. H_0 mean = 100. True mean varies. True σ varies.

Provisional Findings

1) **There is a monotonic relationship between the order of magnitude of the p-value and the relative probability that H_0 is true.**

Provisional Findings

1) There is a monotonic relationship between the order of magnitude of the p-value and the relative probability that H_0 is true.

2) **If using a p-value algorithm to decide whether or not to reject H_0, then (all else being equal):**
   a) **For thick H_0's: (effective α) > (nominal α)**
   b) **For thin H_0's: (effective α) < (nominal α)**

Provisional Findings

1) There is a monotonic relationship between the order of magnitude of the p-value and the relative probability that H_0 is true.

2) If using a p-value algorithm to decide whether or not to reject H_0, then (all else being equal):
   a) For thick H_0's: (effective α) > (nominal α)
   b) For thin H_0's: (effective α) < (nominal α)

3) **Tentatively, these effects seem independent of (a) the size of σ and (b) the method used to obtain the p-values**

A Few References

- Introduction/history of the problem:
  - Ziliak, S.T. and McCloskey, D.N. (2009) The Cult of Statistical Significance. *Proceedings, JSM 2009*
  - Goodman, S.N. (1993) *p* Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate. *American Journal of Epidemiology*. 137(5), 485-496.
- A Bayesian perspective…and re "thickness"
  - Berger, J.O. and Delampady, M. (1987) Testing Precise Hypotheses. *Statistical Science*. 2(3), 317-352.
- "Specified Allowable Error" or "Regions of Indifference" and Tests of Equivalence or Clinical Non-Inferiority
  - Robinson, A.P. and Froese, R.E. (2004) Model Validation Using Equivalence Tests. *Ecological Modeling*. 176, 349-358.

The Problematic Inequality

P([The sample data have the obtained distribution] | [H_0 is true])

Issue 1: "Thickness" of H_0 ?

≠

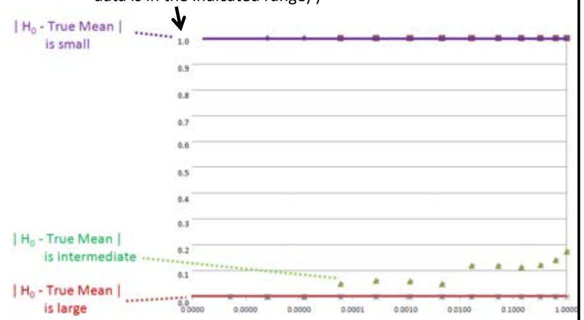
Issue 2: Mismatched structures?

P([H_0 is true] | [The sample data have the obtained distribution])

Is this a true (frequentist) probability?

Pr(the true population mean ≈ H_0 mean | (the p-value returned from the data is in the indicated range))

"Approximately equal to H_0 mean": (All cases: +/- 2 standard errors) 20,000 re-samples



Approximate p-Value based on a t test for μ = (null mean). *n* = 40. H_0 mean = 100. True mean varies. True σ varies.

Provisional Findings

1) Monotonic relationship between … p-value and the relative probability that H_0 is true.

2) a) For thick H_0's: (effective α) > (nominal α)

   b) For thin H_0's: (effective α) < (nominal α)

3) These effects seem independent of (a) size of σ and (b) the method to obtain p-values

4) p-Values *cannot* tell you the "probability that (*on this occasion*) H_0 is true (or false)"

An Additional Challenge?

P([The sample data have the obtained distribution] | [H_0 is true])

Is this really the p-value?

≠

Issue 2: Mismatched structures?

P([H_0 is true] | [The sample data have the obtained distribution])

An Additional Challenge?

*P([*the sample statistic meets {criterion T}*]* **|** *[(H_0 is true)* **and** ({Criterion T} has been pre-determined procedurally from a sample) *]*

≠

Issue 2: Mismatched structures?

P([H_0 is true] **|** *[*the sample statistic meets {criterion T}*])*

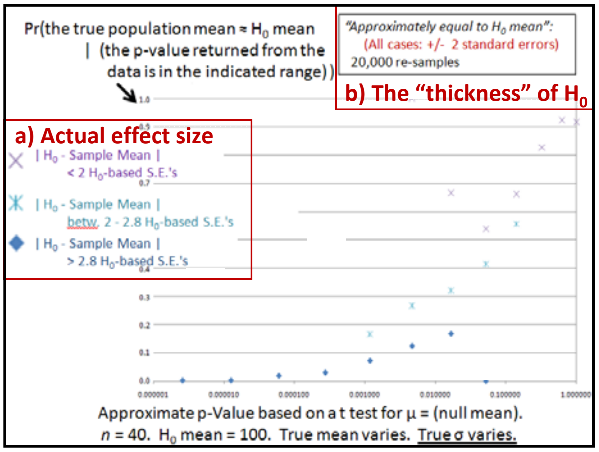
Recommendations

1) Don't give up on *p*-values, but keep clear on what they do—and do not—tell us, and under what conditions.

2) At the very least, provide (or look for) this supplementary information:

a) Actual effect size, *and*

b) The "thickness" of H_0, i.e. the minimum difference that's detectable and/or cared about.


Pr(the true population mean ≈ H0 mean | (the p-value returned from the data is in the indicated range))
"Approximately equal to H0 mean":
(All cases: +/- 2 standard errors)
20,000 re-samples
b) The "thickness" of H0
a) Actual effect size
| H0 - Sample Mean | < 2 H0-based S.E.'s
| H0 - Sample Mean | betw. 2 - 2.8 H0-based S.E.'s
| H0 - Sample Mean | > 2.8 H0-based S.E.'s
1.0
0.7
0.4
0.3
0.2
0.1
0.0
0.000001
0.000010
0.000100
0.001000
0.010000
0.100000
1.000000
Approximate p-Value based on a t test for μ = (null mean).
n = 40. H0 mean = 100. True mean varies. True σ varies.