From “The American Statistician”:

A Proposed Hybrid Effect Size Plus p-Value Criterion: Empirical Evidence Supporting its Use

By William M. Goodman, Susan E. Spruill and Eugene Komaroff (2018)

Reading Course WS 2020/21
November 23, 2020

Outline

1. Introduction
    a. Background
    b. Goals
    c. Definitions
2. Simulation setting
    a. Simulation Setting
    b. Methods
    c. Comparison
    d. Nominal Power
3. Own Simulation
4. Results
5. Limitations
6. Conclusion

Background

Discussions of the use and misuse of p-values

Handling

- ➔ banning the use of null hypothesis significance testing
- ➔ lowering common $p < 0.05$ criterion to $p < 0.005$
- ➔ using alternative statistics, such as confidence intervals, effect size and Bayes factors.

Disagreement

- ➔ relative utility of p-values as evidence for a hypothesis

Agreement

- ➔ p-values are useless if not assessed in the proper context and derived from properly designed studies.

⇨ There are very mixed emotions from the statistical community. This article seeks empirical evidence to help address the issue.

Goals

1. Do p-values have evidential value?

   It is observed that p-values can and do provide evidential information that is relevant for making an inference. But: no stand-alone application.

2. What are the nature and limits of p-values?

3. How do p-values compare with other possible alternative approaches, including a hybrid approach introduced by the authors?

Definitions

MPSD

How close to exactly equal must H0 be to the true parameter to say that H0 is true?

MPSD = minimum practically significant distance

The value that the simulated study's researchers would deem the smallest observed distance from equaling exactly the null that could be considered meaningfully large.

Thick Null

The range of parameter values that would be deemed not meaningful different from equaling the exactly specified point null value.

$$(H_0 - MPSD) \leq \mu \leq (H_0 + MPSD)$$

The **Null Interval** is defined as

$$[(H_0 - MPSD), (H_0 + MPSD)]$$

Simulation Setting

- set-up values randomly generated for each of the four main elements:

  $75 \leq \mu \leq 125$

  $4 \leq \sigma \leq 60$

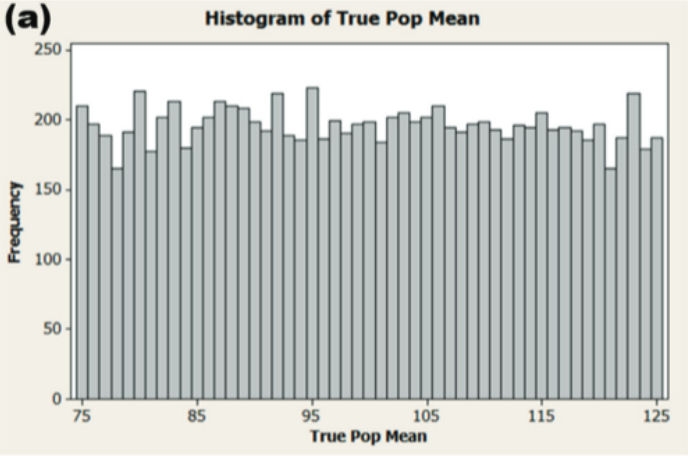
  $5 \leq n \leq 100$

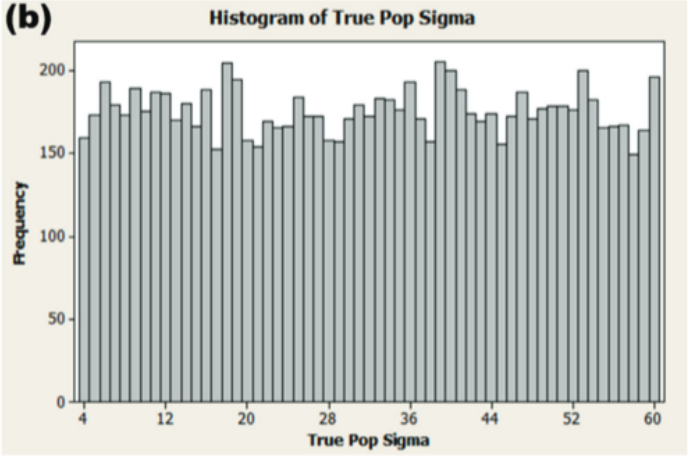
  $2 \leq MPSD \leq 20$

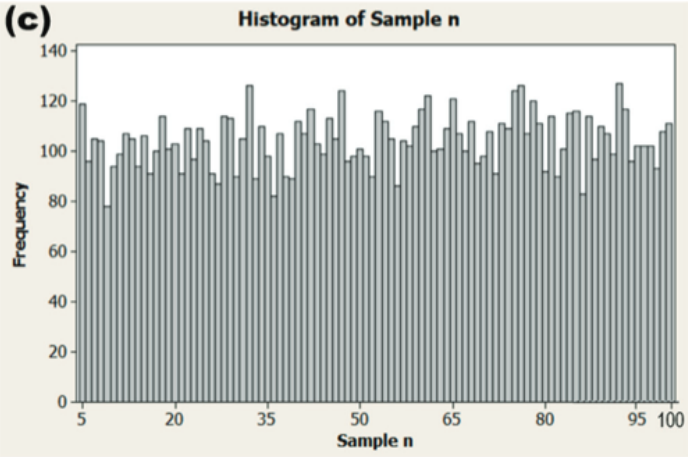
- 10,000 cases
- true μ and σ known (important and in contrast to statistic in practice)

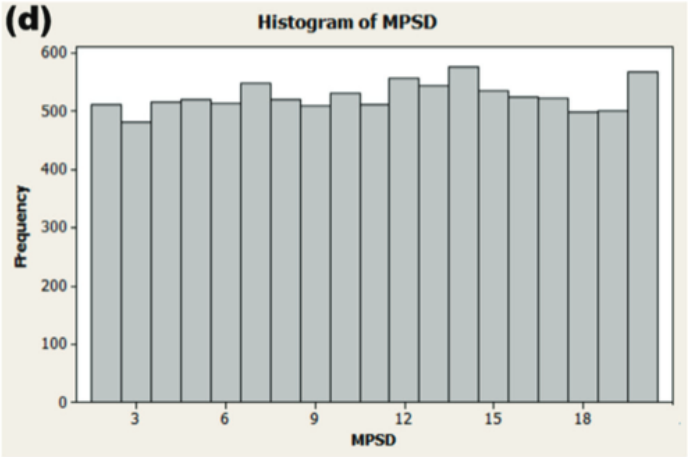
  $H_0 : \mu = 100$

Histogram of sampled set-up values







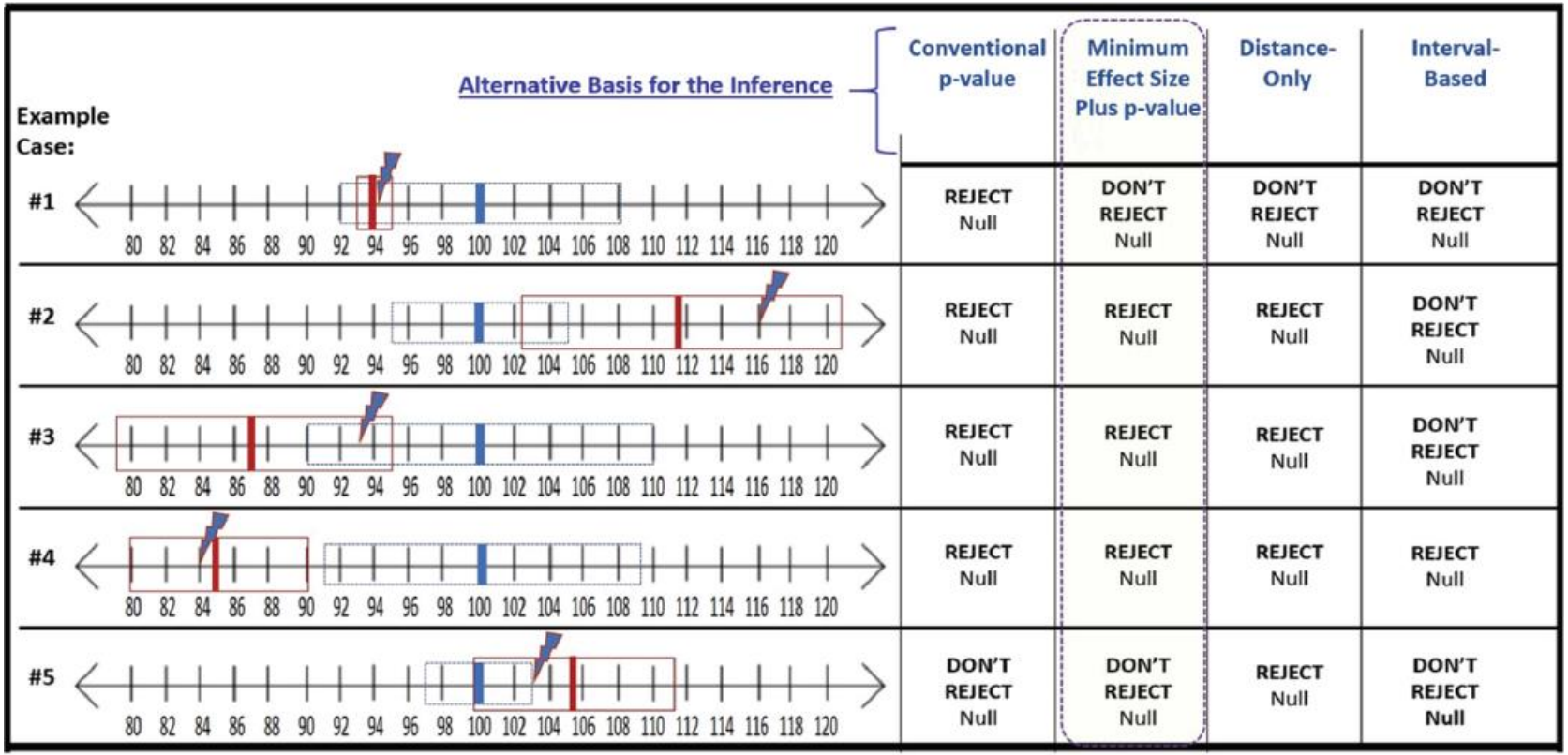


$$75 \leq \mu \leq 125$$
$$4 \leq \sigma \leq 60$$
$$5 \leq n \leq 100$$
$$2 \leq MPSD \leq 20$$

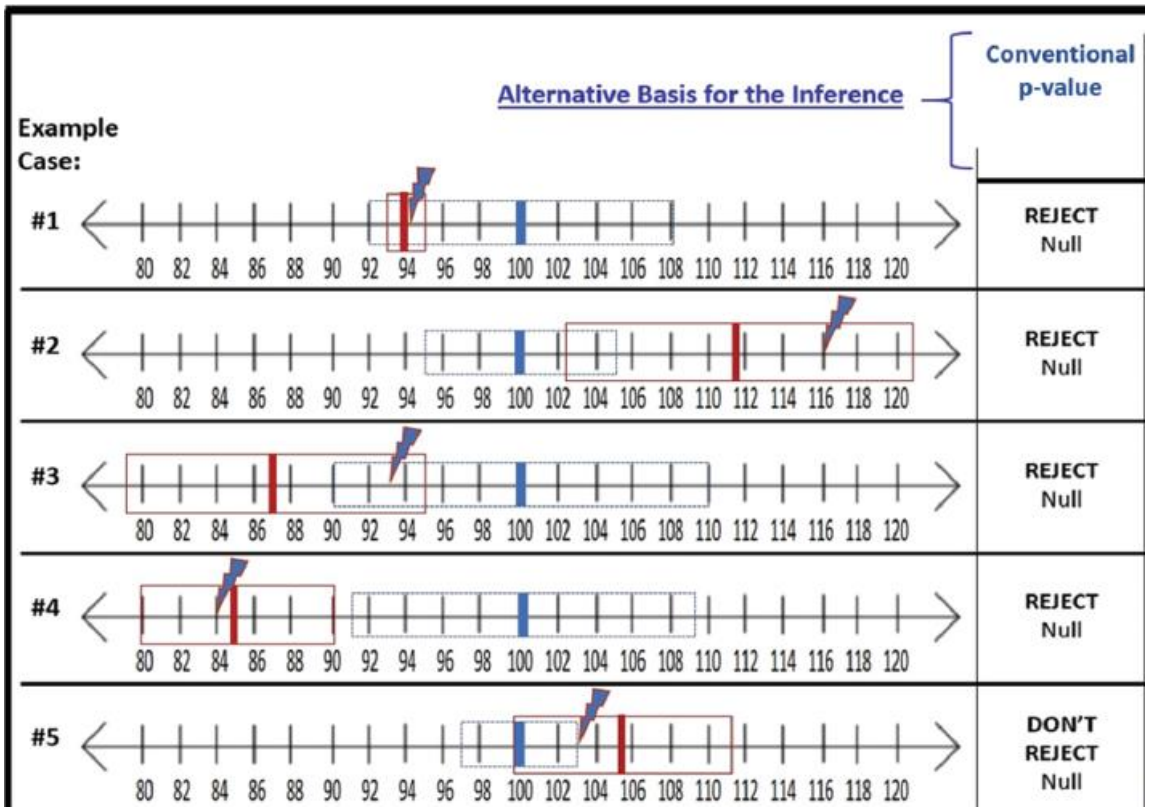
Overview of methods

1. Conventional p-value (/small α)
2. Distance-Only
3. MESP (“Minimum Effect Size Plus p-value”)
4. Interval-Based

Comparison of methods

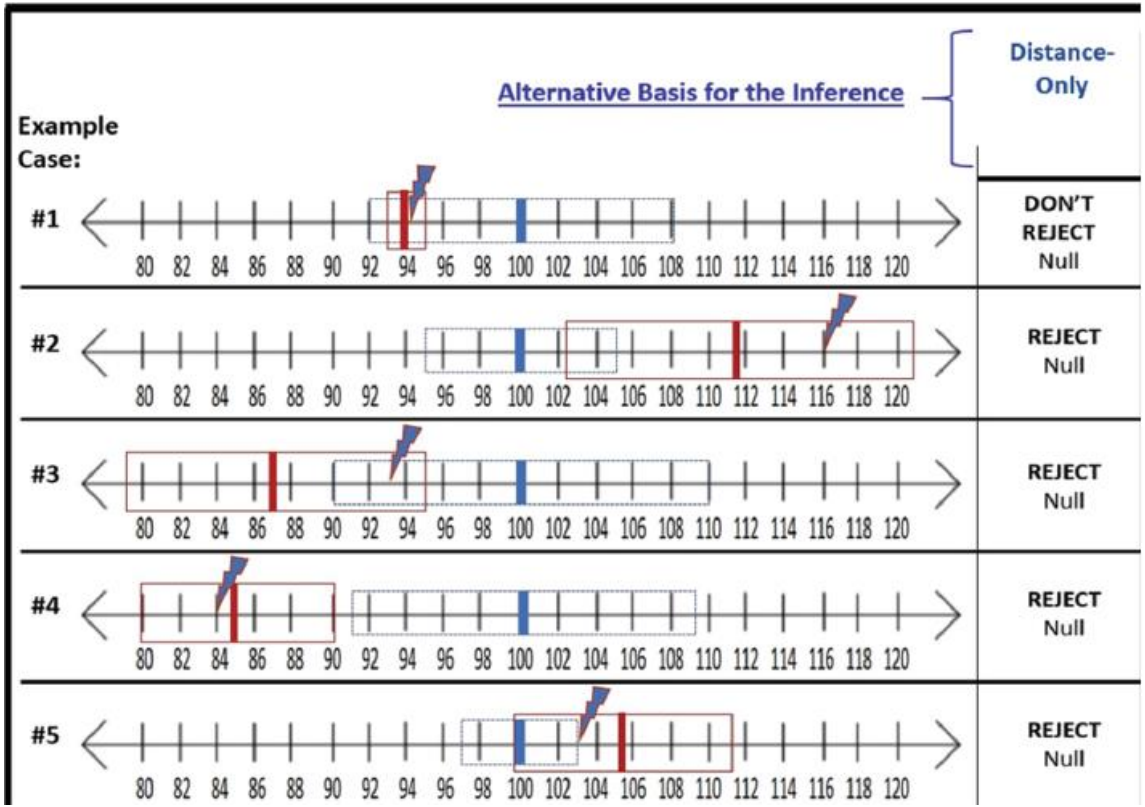


Conventional p-value (/small α)



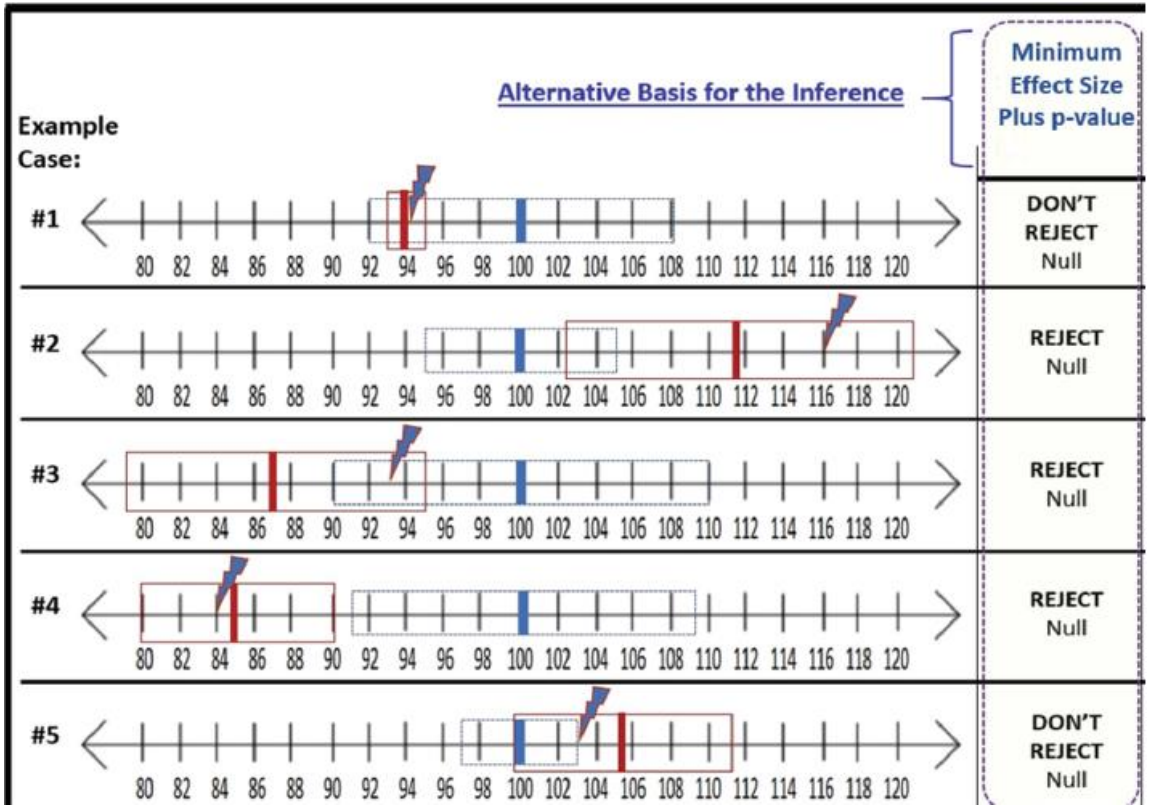
- conventional hypothesis testing with p-value
- colloquially: blue line not in red box
- later in the analysis: same testing but with smaller α (0.005)

Distance-Only



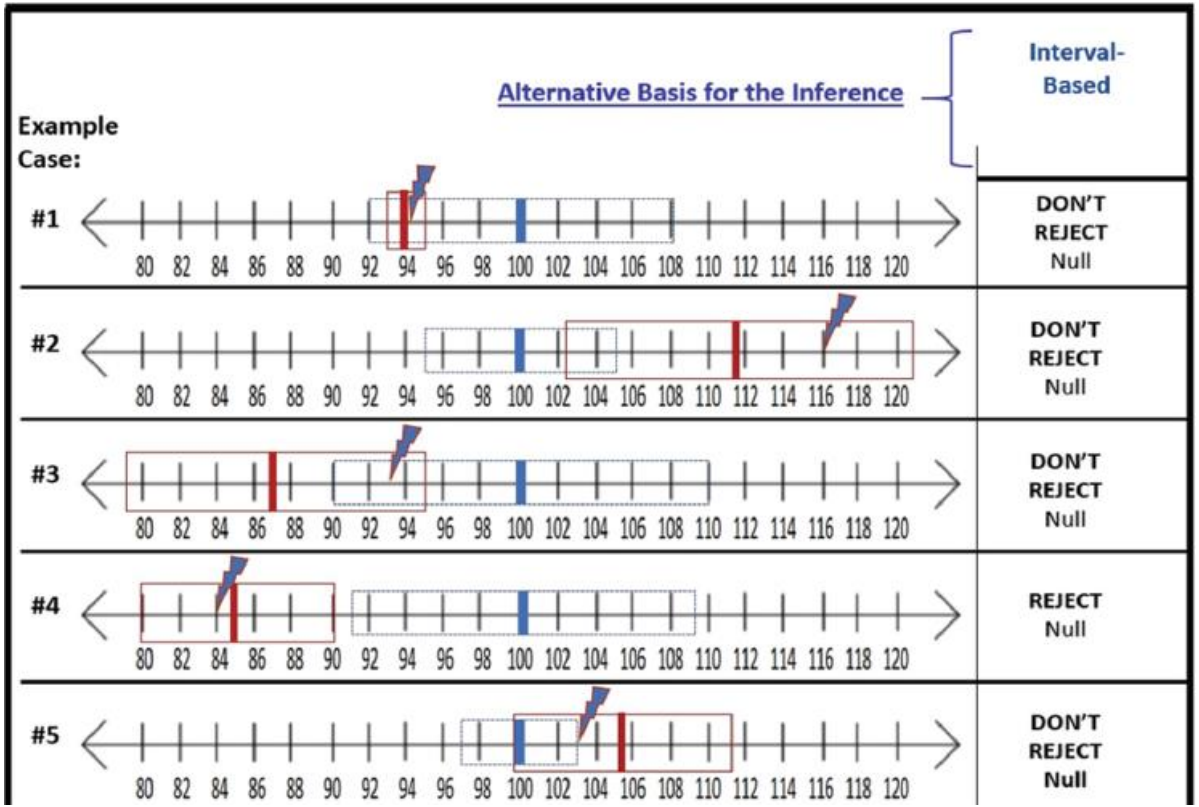
- reject if observed effect size (distance between point null hypothesis and sample mean) greater or equal than MPSD
- Distance only colloquially: red line not in blue box

MESP



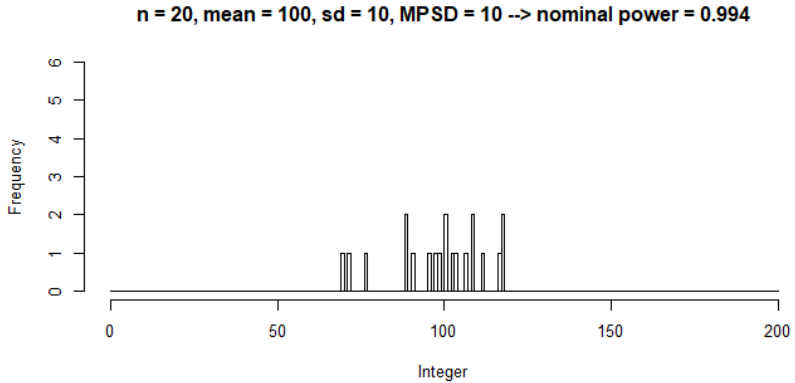
- conditions of conventional p-value and distance-only together → MESP
- colloquially: blue line not in red box **and** red line not in blue box

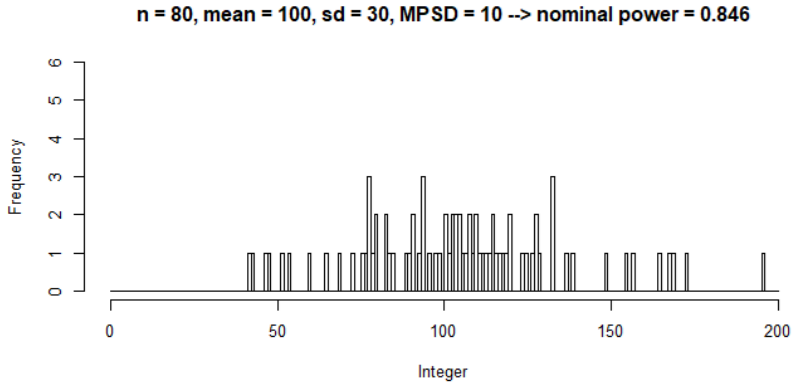
Interval-Based

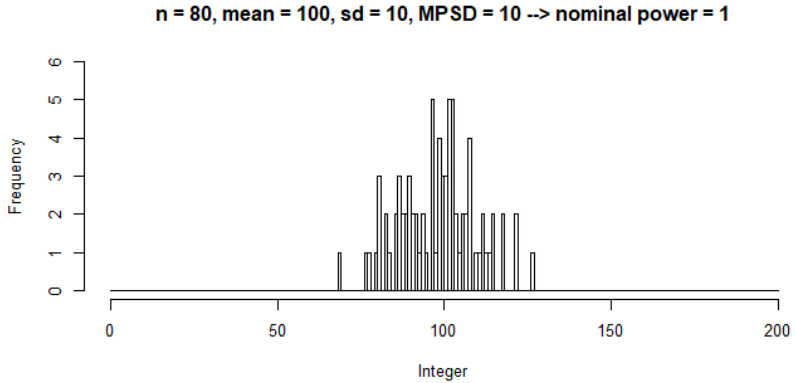


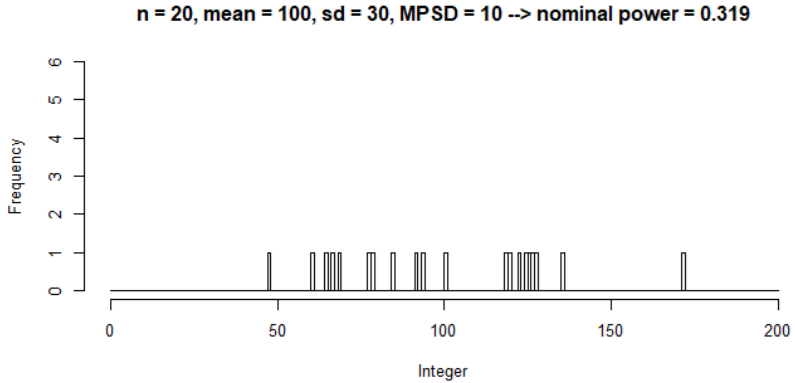
- reject only if thick null and sample-based interval-estimate do not overlap
- demanding method regarding rejection

Nominal Power

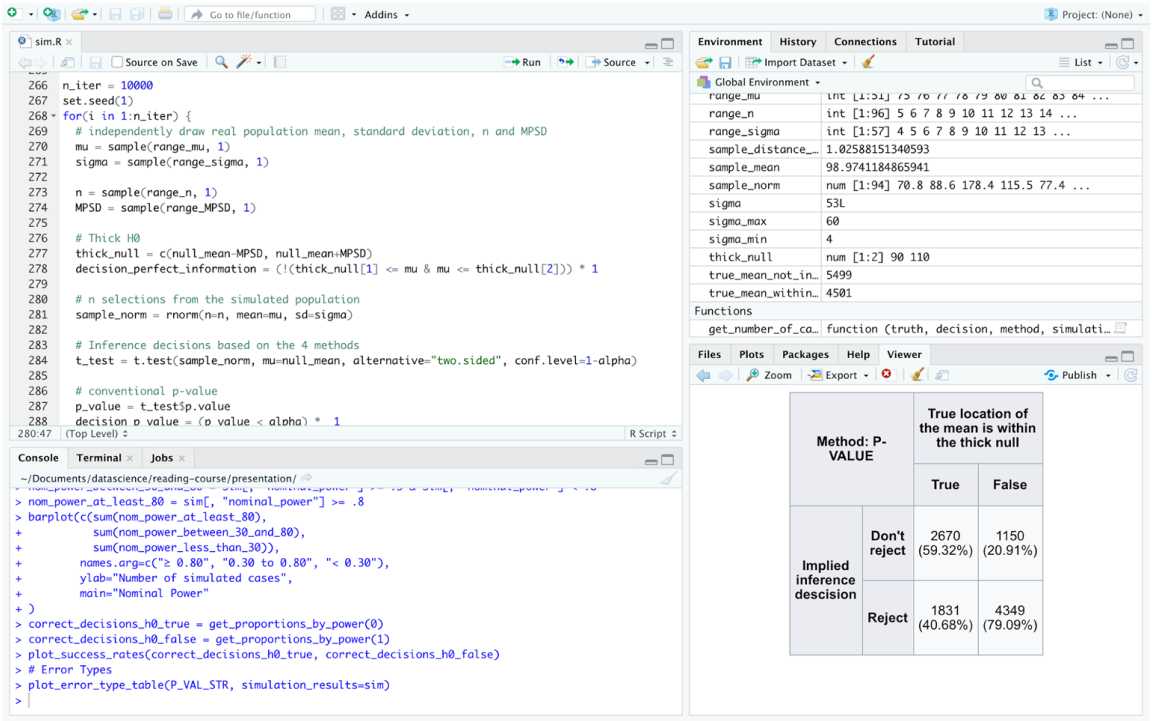








Implementation of the Simulation Approach in R



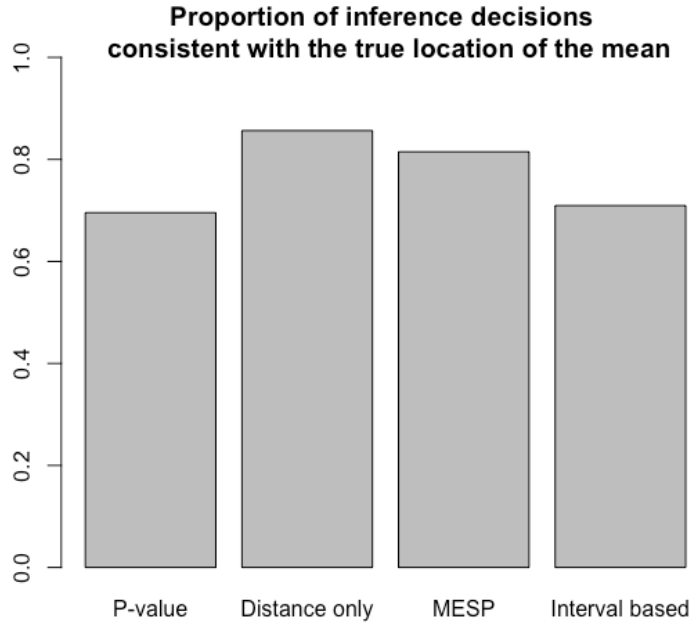
```r
n_iter = 10000
set.seed(1)
for(i in 1:n_iter) {
  # independently draw real population mean, standard deviation, n and MPSD
  mu = sample(range_mu, 1)
  sigma = sample(range_sigma, 1)

  n = sample(range_n, 1)
  MPSD = sample(range_MPSD, 1)

  # Thick H0
  thick_null = c(null_mean-MPSD, null_mean+MPSD)
  decision_perfect_information = (!(thick_null[1] <= mu & mu <= thick_null[2])) * 1

  # n selections from the simulated population
  sample_norm = rnorm(n=n, mean=mu, sd=sigma)

  # Inference decisions based on the 4 methods
  t_test = t.test(sample_norm, mu=null_mean, alternative="two.sided", conf.level=1-alpha)

  # conventional p-value
  p_value = t_test$p.value
  decision_p_value = (p_value < alpha) * 1
```

```
> nom_power_at_least_80 = sim[, "nominal_power"] >= .8
> barplot(c(sum(nom_power_at_least_80),
+           sum(nom_power_between_30_and_80),
+           sum(nom_power_less_than_30)),
+         names.arg=c("≥ 0.80", "0.30 to 0.80", "< 0.30"),
+         ylab="Number of simulated cases",
+         main="Nominal Power"
+ )
> correct_decisions_h0_true = get_proportions_by_power(0)
> correct_decisions_h0_false = get_proportions_by_power(1)
> plot_success_rates(correct_decisions_h0_true, correct_decisions_h0_false)
> # Error Types
> plot_error_type_table(P_VAL_STR, simulation_results=sim)
>
```

		Method: P-VALUE	True location of the mean is within the thick null	
			True	False
Implied inference descision	Don't reject		2670 (59.32%)	1150 (20.91%)
	Reject		1831 (40.68%)	4349 (79.09%)

Results


Does the true location fall
within the bounds of the 'thick null'?
5000
4000
3000
2000
1000
0
YES (4501)
NO (5499)


Proportion of inference decisions
consistent with the true location of the mean
0.0
0.2
0.4
0.6
0.8
1.0
P-value
Distance only
MESP
Interval based

Results

Method: P-VALUE		True location of the mean is within the thick null	
		True	False
Implied inference descision	Don't reject	2680 (59.64%)	1230 (22.34%)
	Reject	1814 (40.36%)	4276 (77.66%)

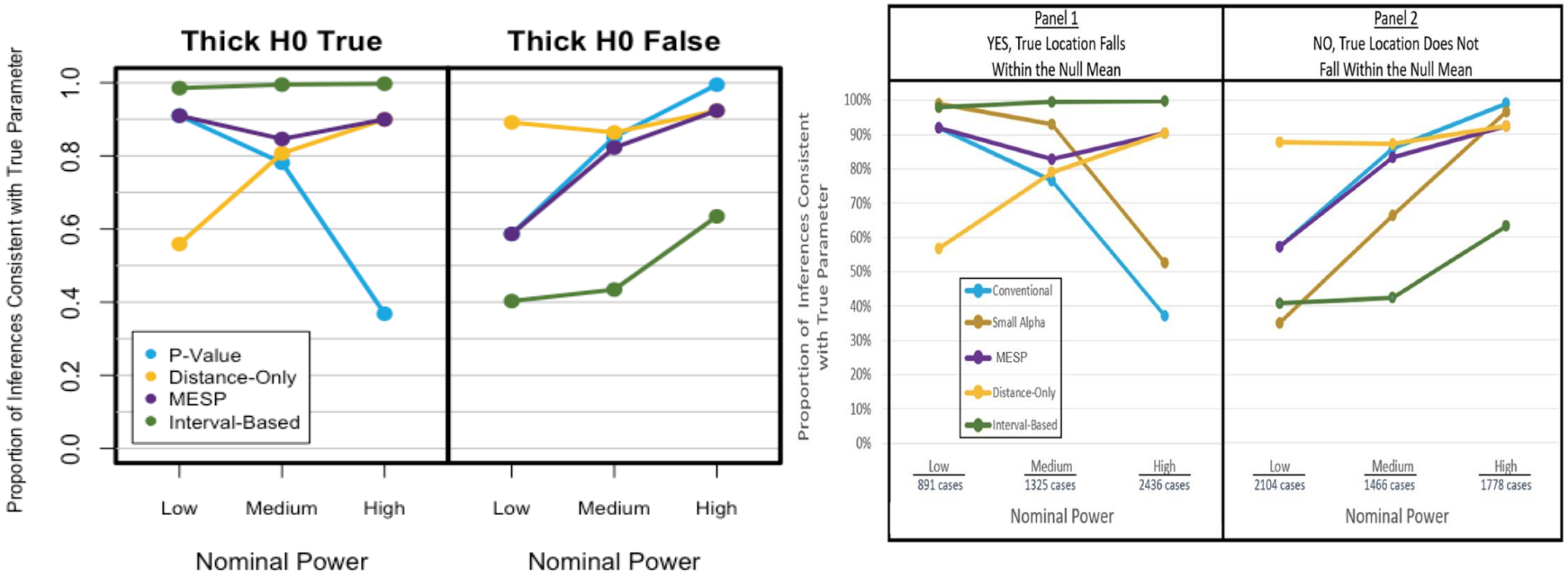
Method: DISTANCE-ONLY		True location of the mean is within the thick null	
		True	False
Implied inference descision	Don't reject	3656 (81.35%)	597 (10.84%)
	Reject	838 (18.65%)	4909 (89.16%)

Results

Method: MESP		True location of the mean is within the thick null	
		True	False
Implied inference descision	Don't reject	4010 (89.23%)	1366 (24.81%)
	Reject	484 (10.77%)	4140 (75.19%)

Method: INTERVAL-BASED		True location of the mean is within the thick null	
		True	False
Implied inference descision	Don't reject	4469 (99.44%)	2880 (52.31%)
	Reject	25 (0.56%)	2626 (47.69%)

Results


Thick H0 True
Thick H0 False
Proportion of Inferences Consistent with True Parameter
0.0
0.2
0.4
0.6
0.8
1.0
P-Value
Distance-Only
MESP
Interval-Based
Low
Medium
High
Nominal Power
Proportion of Inferences Consistent with True Parameter
Panel 1
YES, True Location Falls Within the Null Mean
Panel 2
NO, True Location Does Not Fall Within the Null Mean
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Conventional
Small Alpha
MESP
Distance-Only
Interval-Based
Low
891 cases
Medium
1325 cases
High
2436 cases
Nominal Power
Low
2104 cases
Medium
1466 cases
High
1778 cases
Nominal Power

Limitations

- ➔ *Representativeness of the Model:*
  - ◆ generality of the model has not been tested or demonstrated for nonincluded variations
- ➔ *Generating of Data for Cases:*
  - ◆ the design choices can impact the comparative decision
  - ◆ conclusions always have some relation to the data context
  - ◆ more research is needed for different combinations of scenarios or extremes of distributions
- ➔ *Risks of MPSD “Hacking”:*
  - ◆ "hack" results by setting MPSD sizes that are biased to their own study's advantage (authors appeal to professional ethics)
  - ◆ risks of MESP are not inherently different than those for any inference method

The goal for the simulation was to generate many different combinations of independently generated setup values, to observe the factors’ effects and interactions. The authors have tried to cover as many cases as possible, but there are still some limitations

Conclusion

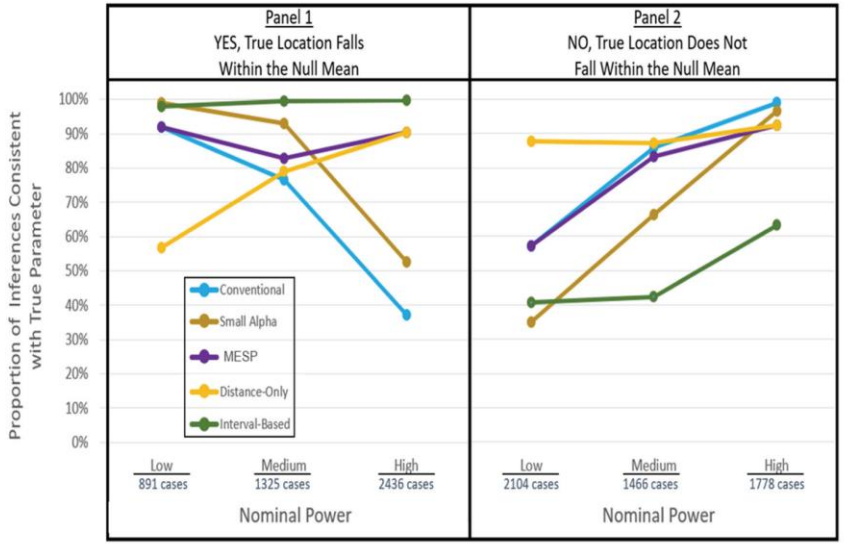


Figure 3. Graph of impact of power, method, and true location of the null on inference success.

- ***interval-based method:*** best method when null is true, worst method when null is false
- ***conventional method:*** good method when null is true and nominal power is low, but very bad method when nominal power is high. If null is false, it is the opposite
- ***small alpha method:*** quite similar to the conventional method
- ***distance only method:*** good method if null is false, not so good method if null is true.
- ***MESP:*** compromise

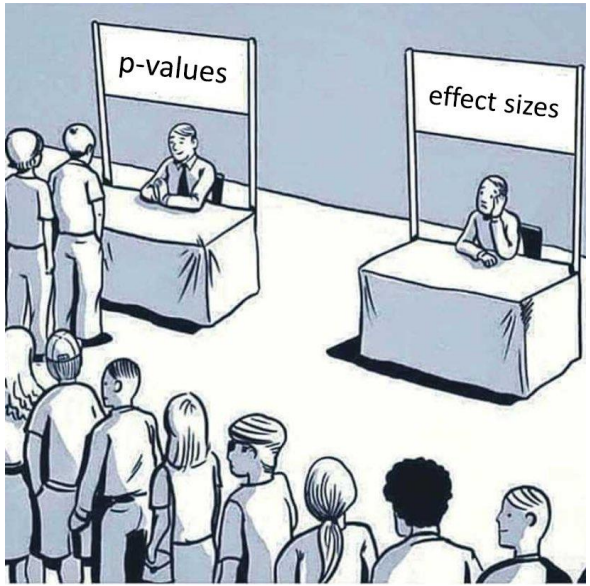
All the compared methods have strengths and weaknesses, and none of them generates an automatic final answer for a definitive inference, based on one application.

Conclusion

- ➔ *desire:* (a) method is not sensitive to factors that are not knowable to the researcher
  (b) method performs well in contexts that the researcher can check for
- ➔ MESP: (+) regardless of whether the unseen real mean happens to be within the thick null or not
  (-) its true power weakens in low nominal power cases, but researcher can respond accordingly
  (+) ready availability of the p-value component of its indicator (not difficult to implement in statistics software)
- ➔ *Own Simulation:* our own simulation agrees with the results of the paper

⇨ NHST model can still have a place in scientific research if results are interpreted properly and not taken as automatically justifying final conclusions. MESP combines the traditional p-value < 0.05 condition with a minimum effect size criterion.

Discussion



Source: William M. Goodman, Susan E. Spruill & Eugene Komaroff (2019) A Proposed Hybrid Effect Size Plus p-Value Criterion: Empirical Evidence Supporting its Use, The American Statistician, 73:sup1, 168-185, DOI: 10.1080/00031305.2018.1564697