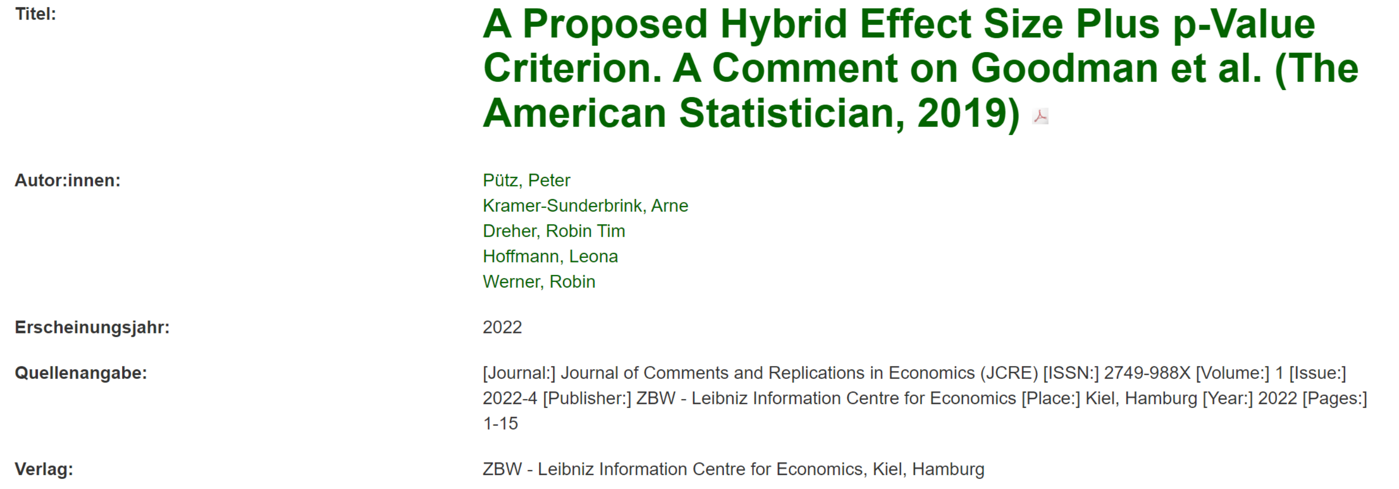
Applied Statistical Analysis II
Replication study

Tianmin Zhang - 23365362

Pütz, P., Kramer-Sunderbrink, A., Dreher, R. T., Hoffmann, L., & Werner, R. (2022). A Proposed Hybrid Effect Size Plus p-Value Criterion. A Comment on Goodman et al.(The American Statistician, 2019). *Journal of Comments and Replications in Economics (JCRE)*, *1*(2022-4), 1-15.



Titel:	**A Proposed Hybrid Effect Size Plus p-Value Criterion. A Comment on Goodman et al. (The American Statistician, 2019)**
Autor:innen:	Pütz, Peter Kramer-Sunderbrink, Arne Dreher, Robin Tim Hoffmann, Leona Werner, Robin
Erscheinungsjahr:	2022
Quellenangabe:	[Journal:] Journal of Comments and Replications in Economics (JCRE) [ISSN:] 2749-988X [Volume:] 1 [Issue:] 2022-4 [Publisher:] ZBW - Leibniz Information Centre for Economics [Place:] Kiel, Hamburg [Year:] 2022 [Pages:] 1-15
Verlag:	ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

Overview

Original

Goodman et al. (2019) introduced a hybrid decision criterion that requires both a statistically significant p-value and a practically significant effect size for an effect to be considered significant. This was proposed as a solution to the commonly criticized practice of relying solely on p-values for determining the significance of study results.

Replication

In the replication and analysis of the study by Goodman et al. (2019), additional covariates and interactions can be considered to potentially enrich the study results: In replication studies, we can introduce theoretically meaningful covariates such as the underlying distribution shape (skewness or kurtosis) of the sample, which may affect the performance of the mixing criteria. The interaction between sample size (n) and minimum practical significant distance (MPSD) can also be explored, as a large sample size may reduce the variability of the effect size estimate and affect the error rate of the decision criteria. In addition, the interaction between effect size and standard deviation may be another covariable to consider, as it may affect the ability to test and the likelihood of rejecting the null hypothesis.

Thick t-test

$$\begin{aligned} p &= P(|\bar{X} - \mu_0| > |\bar{x} - \mu_0| \mid H_0^t) \\ &= \sum_{\dot{\mu}=\mu_0-MPSD}^{\mu_0+MPSD} P(|\bar{X} - \mu_0| > |\bar{x} - \mu_0| \mid \mu = \dot{\mu}) P(\mu = \dot{\mu} \mid H_0^t). \end{aligned}$$

The MESP method and the Thick t-test control errors related to the "thick null hypothesis." MESP combines traditional and distance-based methods, acting like each when appropriate. The Thick t-test, newer and more conservative, rejects the thick null hypothesis less frequently. Both aim to control errors in thick null hypothesis scenarios.

Implementation Steps:

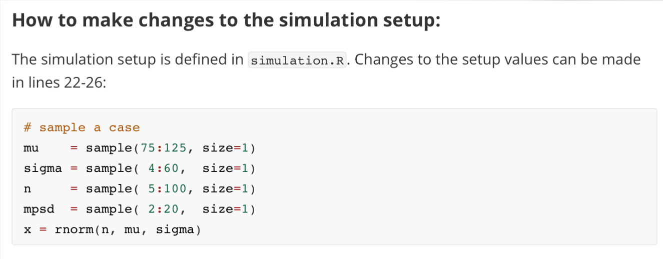
1. Modify simulation.R: Enhance to generate skewness, kurtosis, and interaction terms. Update to accurately reflect distributional characteristics and capture variable interactions.

2. Update methods.R: Refine decision functions to incorporate new covariates and interactions. Adjust criteria to consider distributional characteristics, utilizing additional information for informed decisions.

3. Adjust analysis tools.R: Adapt analysis functions to evaluate results with respect to new covariates and interaction terms. Update metrics calculation considering effects on the hybrid decision criterion's performance.

4. Run updated simulation results.R: Execute to obtain new results reflecting the impact of added covariates and interactions. Validate outputs align with expected outcomes based on modified model and decision functions.

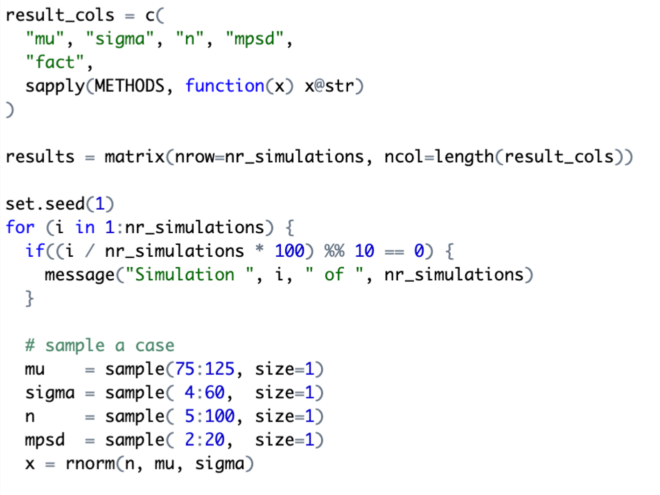
5. Interpretation and Discussion: Analyze results, discuss implications of added covariates, and identify scenarios where criterion performs differently.



How to make changes to the simulation setup:

The simulation setup is defined in `simulation.R`. Changes to the setup values can be made in lines 22-26:

```
# sample a case
mu    = sample(75:125, size=1)
sigma = sample( 4:60,  size=1)
n     = sample( 5:100, size=1)
mpsd  = sample( 2:20,  size=1)
x = rnorm(n, mu, sigma)
```

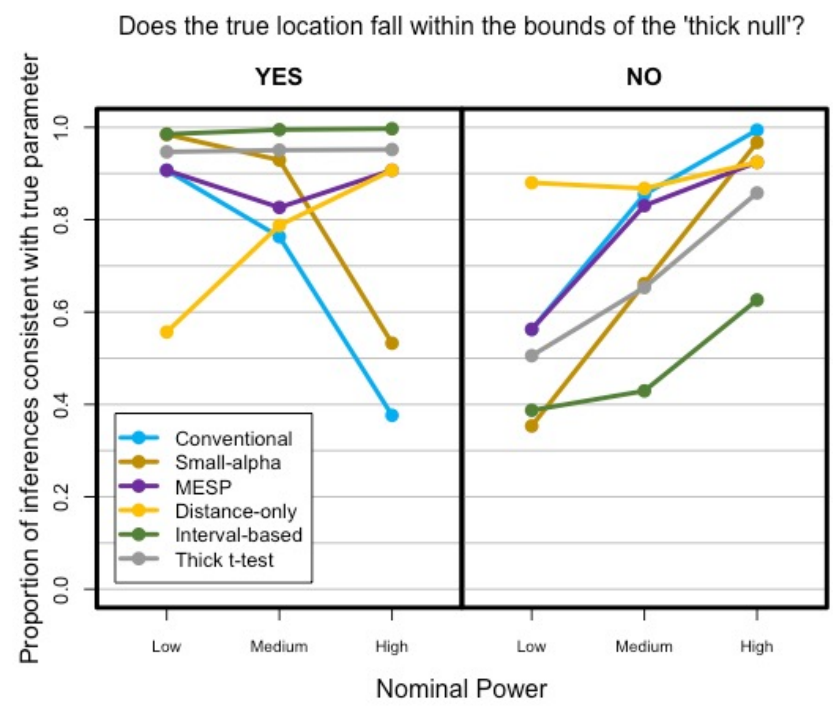


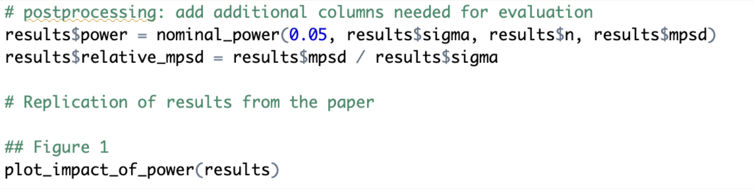
```
result_cols = c(
  "mu", "sigma", "n", "mpsd",
  "fact",
  sapply(METHODS, function(x) x@str)
)

results = matrix(nrow=nr_simulations, ncol=length(result_cols))

set.seed(1)
for (i in 1:nr_simulations) {
  if((i / nr_simulations * 100) %% 10 == 0) {
    message("Simulation ", i, " of ", nr_simulations)
  }

  # sample a case
  mu    = sample(75:125, size=1)
  sigma = sample( 4:60,  size=1)
  n     = sample( 5:100, size=1)
  mpsd  = sample( 2:20,  size=1)
  x = rnorm(n, mu, sigma)
```

Replication

Figure 1:

Graph Purpose: Illustrates the success rate of inference for different decision methods across varying nominal power, true null hypothesis location, and consistency with true parameter range.





```
# postprocessing: add additional columns needed for evaluation
results$power = nominal_power(0.05, results$sigma, results$n, results$mpsd)
results$relative_mpsd = results$mpsd / results$sigma

# Replication of results from the paper

## Figure 1
plot_impact_of_power(results)
```

X-axis: Nominal Power (represents the set power level of the statistical test).

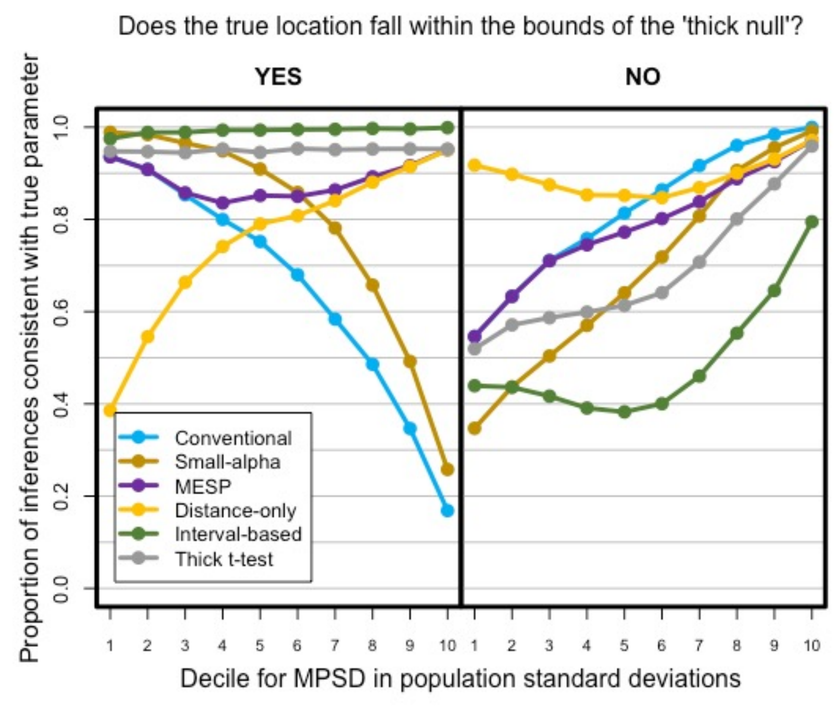
Y-axis: Proportion of inferences consistent with true parameter (indicating the success rate of inference).

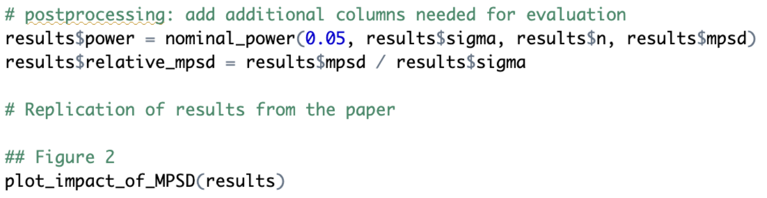
Five Lines:

1. Conventional: Traditional t-test method.
2. Small-alpha Variant: Variant of t-test method with reduced α level to 0.005.
3. MESP: Inference using MESP method.
4. Distance-only: Inference using distance-based method.
5. Interval-based Thick t-test: Inference using interval-based thick t-test.

Figure 2:

Graph Objective: Impact of relative MPSD and method on inference success.
Demonstrates the inference success rate of different decision methods within varying ranges of relative Minimum Detectable Effect (MPSD).





```
# postprocessing: add additional columns needed for evaluation
results$power = nominal_power(0.05, results$sigma, results$n, results$mpsd)
results$relative_mpsd = results$mpsd / results$sigma

# Replication of results from the paper

## Figure 2
plot_impact_of_MPSD(results)
```

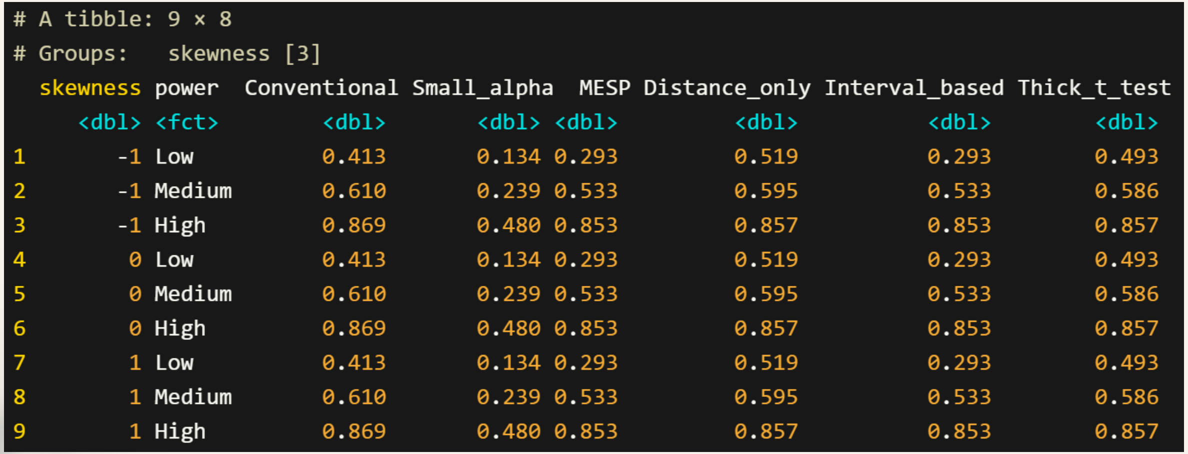
X-axis: Decile for MPSD in population standard deviations. Represents the range of relative Minimum Detectable Effect (MPSD).
Y-axis: Proportion of inferences consistent with true parameter. Indicates the proportion of inference results consistent with the true parameter, i.e., inference success rate.

Five Lines:

1. Conventional: Traditional t-test method.
2. Small-alpha Variant: Variant of t-test method with reduced α level to 0.005.
3. MESP: Inference using MESP method.
4. Distance-only: Inference using distance-based method.
5. Interval-based Thick t-test: Inference using interval-based thick t-test.

Twist

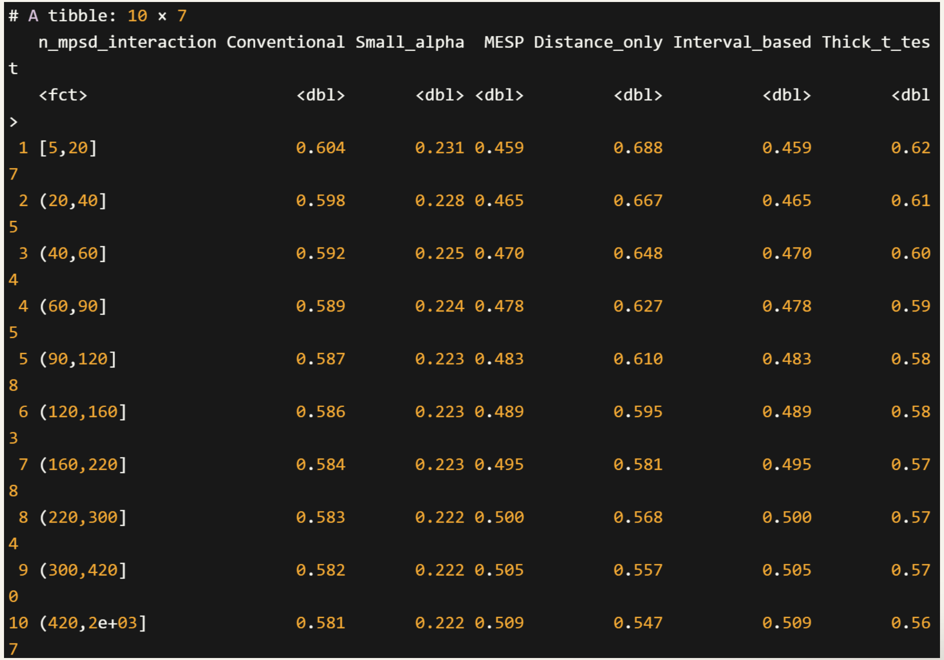
Impact of Power and Skewness



```
# A tibble: 9 × 8
# Groups:   skewness [3]
  skewness power  Conventional Small_alpha  MESP Distance_only Interval_based Thick_t_test
     <dbl> <fct>         <dbl>       <dbl> <dbl>         <dbl>          <dbl>        <dbl>
1       -1 Low           0.413       0.134 0.293         0.519          0.293        0.493
2       -1 Medium        0.610       0.239 0.533         0.595          0.533        0.586
3       -1 High          0.869       0.480 0.853         0.857          0.853        0.857
4        0 Low           0.413       0.134 0.293         0.519          0.293        0.493
5        0 Medium        0.610       0.239 0.533         0.595          0.533        0.586
6        0 High          0.869       0.480 0.853         0.857          0.853        0.857
7        1 Low           0.413       0.134 0.293         0.519          0.293        0.493
8        1 Medium        0.610       0.239 0.533         0.595          0.533        0.586
9        1 High          0.869       0.480 0.853         0.857          0.853        0.857
```

The impact of power and skewness analysis shows that the performance of all methods improves as power increases, regardless of the skewness level. The hybrid criterion (MESP and Interval-based) maintains an advantage over the conventional t-test across all power and skewness levels.

Impact of n*MPSD Interaction:



```
# A tibble: 10 × 7
   n_mpsd_interaction Conventional Small_alpha  MESP Distance_only Interval_based Thick_t_tes
t
   <fct>                     <dbl>       <dbl> <dbl>         <dbl>          <dbl>       <dbl
>
 1 [5,20]                    0.604       0.231 0.459         0.688          0.459       0.62
7
 2 (20,40]                   0.598       0.228 0.465         0.667          0.465       0.61
5
 3 (40,60]                   0.592       0.225 0.470         0.648          0.470       0.60
4
 4 (60,90]                   0.589       0.224 0.478         0.627          0.478       0.59
5
 5 (90,120]                  0.587       0.223 0.483         0.610          0.483       0.58
8
 6 (120,160]                 0.586       0.223 0.489         0.595          0.489       0.58
3
 7 (160,220]                 0.584       0.223 0.495         0.581          0.495       0.57
8
 8 (220,300]                 0.583       0.222 0.500         0.568          0.500       0.57
4
 9 (300,420]                 0.582       0.222 0.505         0.557          0.505       0.57
0
10 (420,2e+03]               0.581       0.222 0.509         0.547          0.509       0.56
7
```

The impact of the n*MPSD interaction reveals that as the interaction term increases (i.e., larger sample sizes and/or larger MPSD), the performance of all methods decreases slightly. However, the hybrid criterion still outperforms the conventional t-test for all levels of the interaction term.

These results provide a more nuanced understanding of the hybrid criterion's performance under different conditions, but overall, they still support the original findings of Goodman et al. (2019) regarding the advantages of the hybrid approach.

My contribution:

1. Richness and comprehensiveness: By adding additional covariables and interactions, I made the research model more comprehensive and able to consider more factors that may affect the results, thus providing a more comprehensive understanding of the research phenomenon.

2. Deepening understanding: By expanding the model, I deepened my understanding of the research phenomenon. Additional covariates and interactions provide more information, helping to reveal the mechanisms and influencing factors behind the studied phenomena.

3. Improved explanatory power: Newly added variables and interaction terms can help explain the model results and make the research conclusions more credible and convincing. This can enhance the interpretability and reproducibility of the study.

4. Expand the research field: By introducing new variables and interactions, I bring new perspectives and exploration directions to the research field. This helps to drive research progress in the field and may trigger more interest and exploration of related research.

5. Enhanced applicability: The expansion of the model can make it more in line with the actual application scenario, thus improving the practicability and applicability of the research results. This helps translate research findings into practical applications or policy recommendations.

Thanks