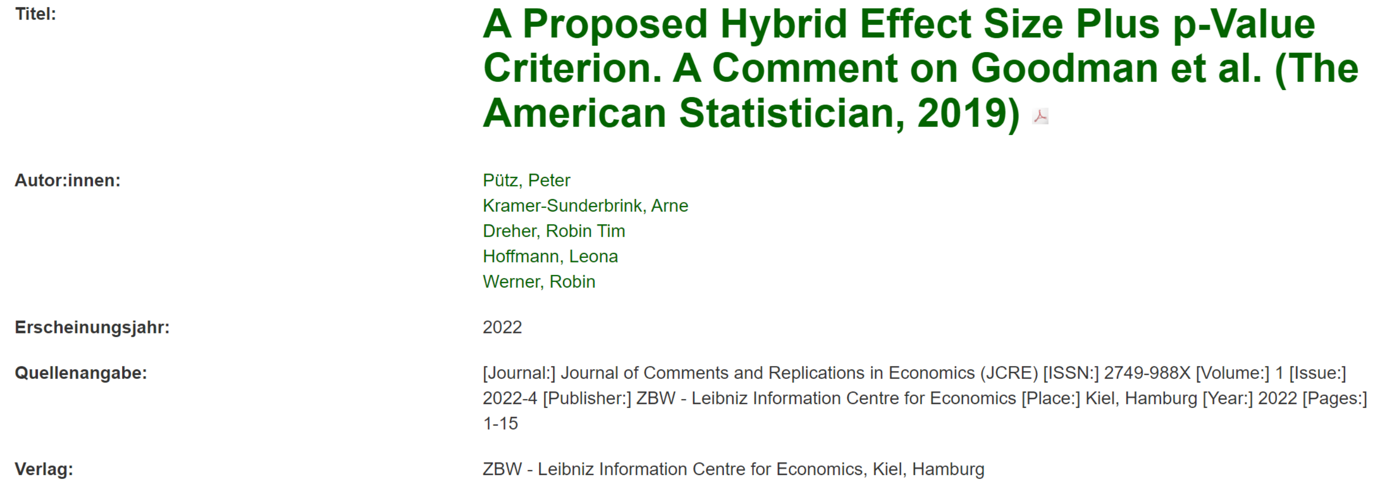
Applied Statistical Analysis II

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

Background

Title: A Proposed Hybrid Effect Size Plus p-Value Criterion. A Comment on Goodman et al.
Authors: Pütz, Peter; Kramer-Sunderbrink, Arne; Dreher, Robin Tim; Hoffmann, Leona; Werner, Robin
Source: Journal of Comments and Replications in Economics (JCRE)
Date: 2022
Link: doi.org/10.18718/81781.26
Version: hdl.handle.net/10419/267164

The paper, titled "A Proposed Hybrid Effect Size Plus p-Value Criterion," authored by Pütz et al., was published in the Journal of Comments and Replications in Economics (JCRE) in 2022. It serves as a commentary and response to an article by Goodman et al. published in The American Statistician in 2019.

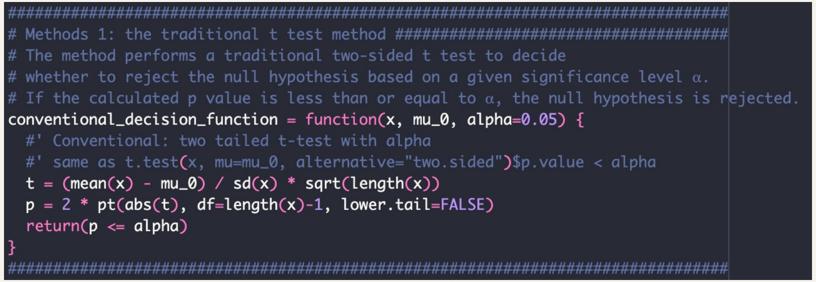
Overview

Original
Goodmp-value and a practically an et al. (2019) introduced a hybrid decision criterion that requires both a statistically significant 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.

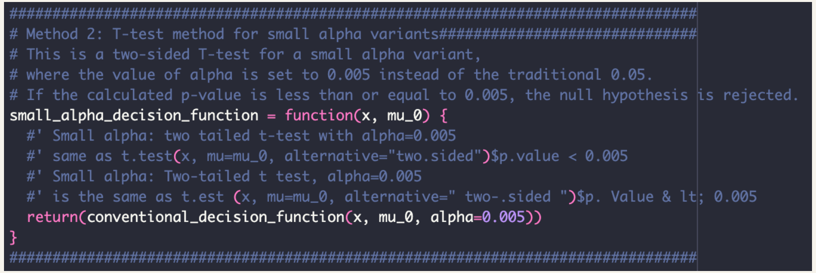
Simulation research methods：

In this study, the authors employed simulation research methods designed to evaluate the performance of various statistical analysis methods in different situations.



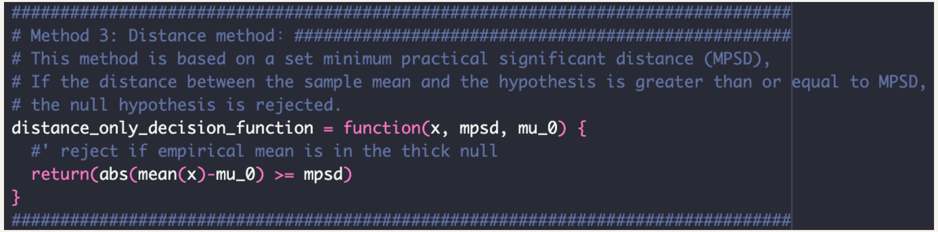
```
################################################################################
# Methods 1: the traditional t test method #####################################
# The method performs a traditional two-sided t test to decide
# whether to reject the null hypothesis based on a given significance level α.
# If the calculated p value is less than or equal to α, the null hypothesis is rejected.
conventional_decision_function = function(x, mu_0, alpha=0.05) {
  #' Conventional: two tailed t-test with alpha
  #' same as t.test(x, mu=mu_0, alternative="two.sided")$p.value < alpha
  t = (mean(x) - mu_0) / sd(x) * sqrt(length(x))
  p = 2 * pt(abs(t), df=length(x)-1, lower.tail=FALSE)
  return(p <= alpha)
}
################################################################################
```

Traditional T-test method: The traditional two-sided T-test is used to decide whether to reject the null hypothesis, based on a given significance level α. If the calculated p-value is less than or equal to α, the null hypothesis is rejected.



```
################################################################################
# Method 2: T-test method for small alpha variants##############################
# This is a two-sided T-test for a small alpha variant,
# where the value of alpha is set to 0.005 instead of the traditional 0.05.
# If the calculated p-value is less than or equal to 0.005, the null hypothesis is rejected.
small_alpha_decision_function = function(x, mu_0) {
  #' Small alpha: two tailed t-test with alpha=0.005
  #' same as t.test(x, mu=mu_0, alternative="two.sided")$p.value < 0.005
  #' Small alpha: Two-tailed t test, alpha=0.005
  #' is the same as t.est (x, mu=mu_0, alternative=" two-.sided ")$p. Value & lt; 0.005
  return(conventional_decision_function(x, mu_0, alpha=0.005))
}
################################################################################
```

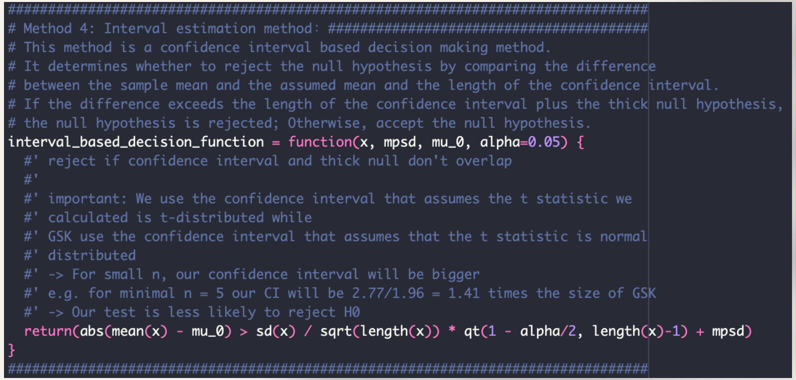
T-test method for the small alpha variant: This is a two-sided T-test where the value of alpha is set to 0.005 instead of the traditional 0.05. If the calculated p-value is less than or equal to 0.005, the null hypothesis is rejected.



```r
################################################################################
# Method 3: Distance method: ###################################################
# This method is based on a set minimum practical significant distance (MPSD),
# If the distance between the sample mean and the hypothesis is greater than or equal to MPSD,
# the null hypothesis is rejected.
distance_only_decision_function = function(x, mpsd, mu_0) {
  #' reject if empirical mean is in the thick null
  return(abs(mean(x)-mu_0) >= mpsd)
}
################################################################################
```

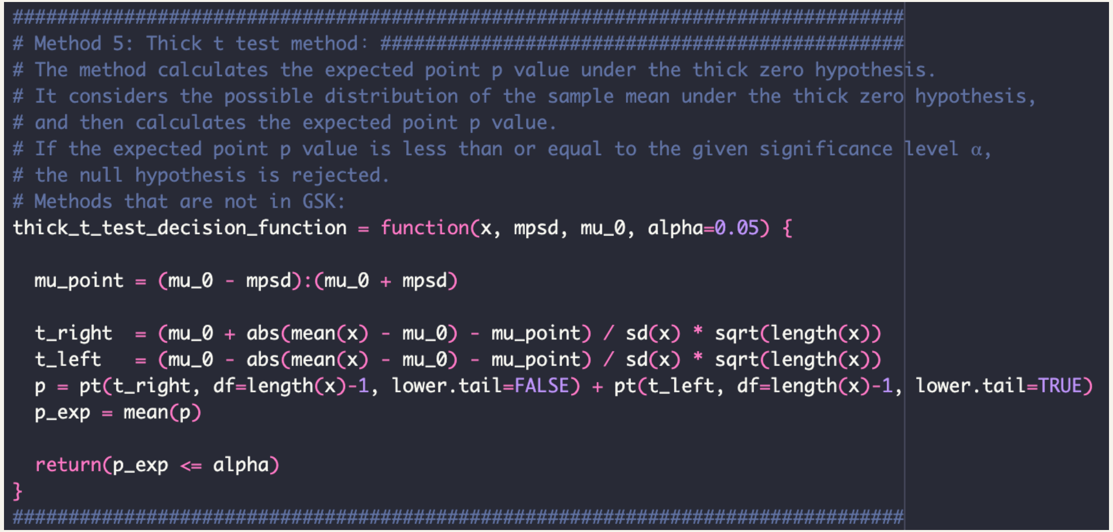
Distance method: Based on a set minimum practical significant distance (MPSD), the null hypothesis is rejected if the distance between the sample mean and the assumed mean is greater than or equal to the MPSD.

Interval estimation method: This is a decision method based on confidence intervals. It determines whether to reject the null hypothesis by comparing the difference between the sample mean and the assumed mean and the length of the confidence interval. If the difference exceeds the length of the confidence interval plus the thick null hypothesis, the null hypothesis is rejected. Otherwise, accept the null hypothesis.



```r
################################################################################
# Method 4: Interval estimation method: ########################################
# This method is a confidence interval based decision making method.
# It determines whether to reject the null hypothesis by comparing the difference
# between the sample mean and the assumed mean and the length of the confidence interval.
# If the difference exceeds the length of the confidence interval plus the thick null hypothesis,
# the null hypothesis is rejected; Otherwise, accept the null hypothesis.
interval_based_decision_function = function(x, mpsd, mu_0, alpha=0.05) {
  #' reject if confidence interval and thick null don't overlap
  #'
  #' important: We use the confidence interval that assumes the t statistic we
  #' calculated is t-distributed while
  #' GSK use the confidence interval that assumes that the t statistic is normal
  #' distributed
  #' -> For small n, our confidence interval will be bigger
  #' e.g. for minimal n = 5 our CI will be 2.77/1.96 = 1.41 times the size of GSK
  #' -> Our test is less likely to reject H0
  return(abs(mean(x) - mu_0) > sd(x) / sqrt(length(x)) * qt(1 - alpha/2, length(x)-1) + mpsd)
}
################################################################################
```

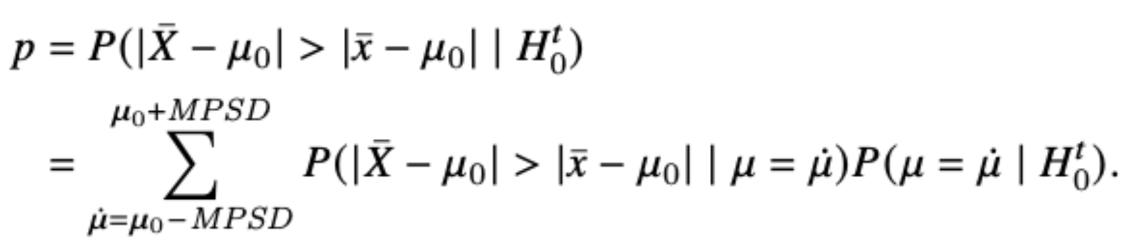
Thick t-test



```
################################################################################
# Method 5: Thick t test method: ###############################################
# The method calculates the expected point p value under the thick zero hypothesis.
# It considers the possible distribution of the sample mean under the thick zero hypothesis,
# and then calculates the expected point p value.
# If the expected point p value is less than or equal to the given significance level α,
# the null hypothesis is rejected.
# Methods that are not in GSK:
thick_t_test_decision_function = function(x, mpsd, mu_0, alpha=0.05) {

  mu_point = (mu_0 - mpsd):(mu_0 + mpsd)

  t_right  = (mu_0 + abs(mean(x) - mu_0) - mu_point) / sd(x) * sqrt(length(x))
  t_left   = (mu_0 - abs(mean(x) - mu_0) - mu_point) / sd(x) * sqrt(length(x))
  p = pt(t_right, df=length(x)-1, lower.tail=FALSE) + pt(t_left, df=length(x)-1, lower.tail=TRUE)
  p_exp = mean(p)

  return(p_exp <= alpha)
}
################################################################################
```

Thickness t test method: This method calculates the expected point P-value under the thickness zero hypothesis, considers the possible distribution of the sample mean under the thickness zero hypothesis, and calculates the expected point P-value. If the expected point p value is less than or equal to the given significance level α, the null hypothesis is rejected.

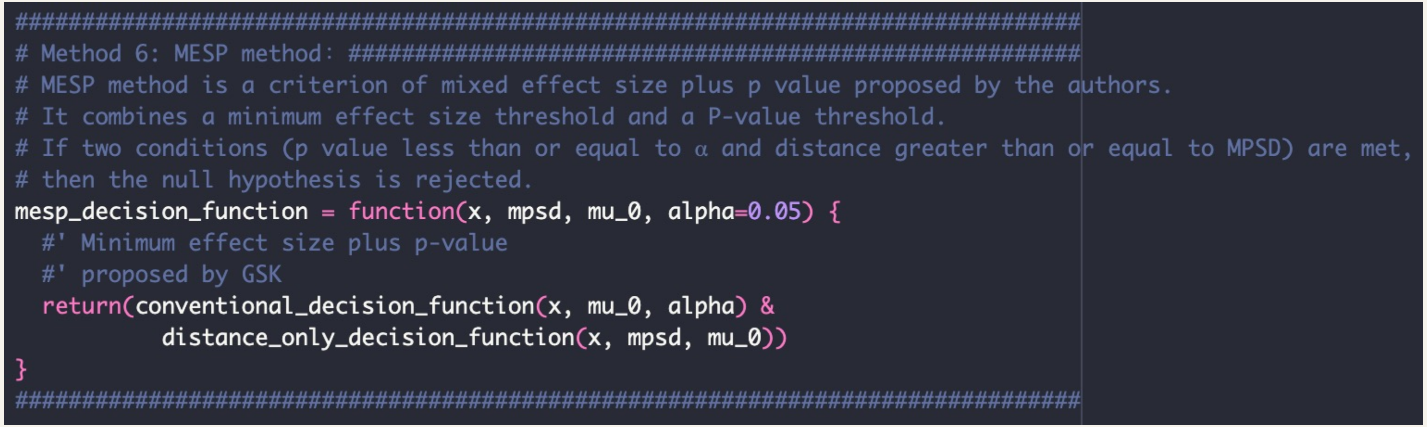
Thick t test method formula：



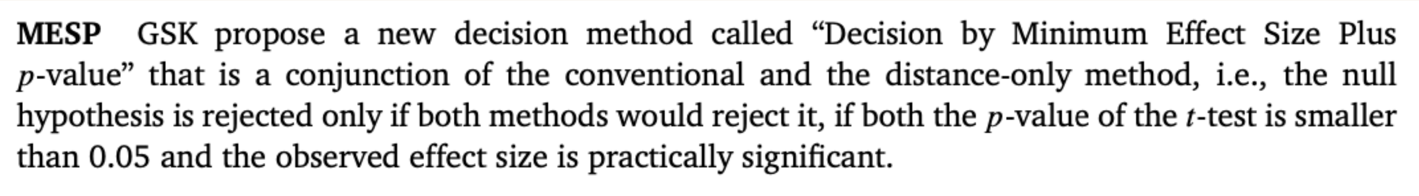
$$\begin{aligned} p &= P(|\bar{X} - \mu_0| > |\bar{x} - \mu_0| \mid H_0^t) \\ &= \sum_{\dot{\mu}=\mu_0-MPSD}^{\mu_0+MPSD} P(|\bar{X} - \mu_0| > |\bar{x} - \mu_0| \mid \mu = \dot{\mu}) P(\mu = \dot{\mu} \mid H_0^t). \end{aligned}$$

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.

MESP :



```r
################################################################################
# Method 6: MESP method: #######################################################
# MESP method is a criterion of mixed effect size plus p value proposed by the authors.
# It combines a minimum effect size threshold and a P-value threshold.
# If two conditions (p value less than or equal to α and distance greater than or equal to MPSD) are met,
# then the null hypothesis is rejected.
mesp_decision_function = function(x, mpsd, mu_0, alpha=0.05) {
  #' Minimum effect size plus p-value
  #' proposed by GSK
  return(conventional_decision_function(x, mu_0, alpha) &
           distance_only_decision_function(x, mpsd, mu_0))
}
################################################################################
```



MESP GSK propose a new decision method called "Decision by Minimum Effect Size Plus *p*-value" that is a conjunction of the conventional and the distance-only method, i.e., the null hypothesis is rejected only if both methods would reject it, if both the *p*-value of the *t*-test is smaller than 0.05 and the observed effect size is practically significant.

Compare the performance of different decision methods:

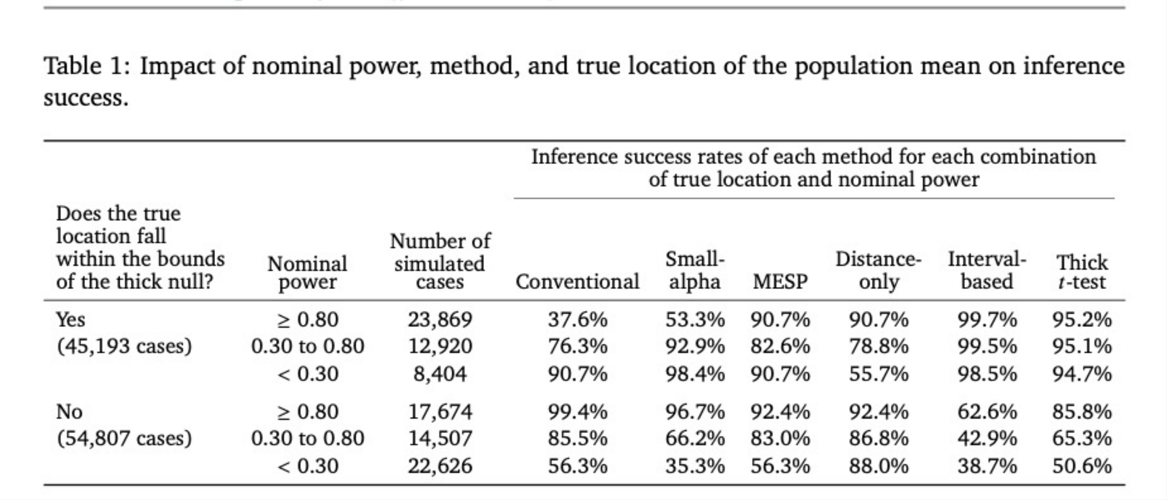


Table 1: Impact of nominal power, method, and true location of the population mean on inference success.

			Inference success rates of each method for each combination of true location and nominal power					
Does the true location fall within the bounds of the thick null?	Nominal power	Number of simulated cases	Conventional	Small-alpha	MESP	Distance-only	Interval-based	Thick *t*-test
Yes	≥ 0.80	23,869	37.6%	53.3%	90.7%	90.7%	99.7%	95.2%
(45,193 cases)	0.30 to 0.80	12,920	76.3%	92.9%	82.6%	78.8%	99.5%	95.1%
	< 0.30	8,404	90.7%	98.4%	90.7%	55.7%	98.5%	94.7%
No	≥ 0.80	17,674	99.4%	96.7%	92.4%	92.4%	62.6%	85.8%
(54,807 cases)	0.30 to 0.80	14,507	85.5%	66.2%	83.0%	86.8%	42.9%	65.3%
	< 0.30	22,626	56.3%	35.3%	56.3%	88.0%	38.7%	50.6%

Table 1: compares the inference success rates of different methods under various conditions. Factors investigated include the true parameter position, the method used, and the nominal power level. Results indicate that certain methods perform better when the true parameter falls within the thick null hypothesis interval, with nominal power level appearing to have little influence. However, when the true parameter lies outside this interval, method selection significantly impacts success rates, highlighting the importance of choosing the appropriate statistical inference method for accurate results. Overall, the table provides valuable insights for researchers in selecting suitable methods for inference analysis.

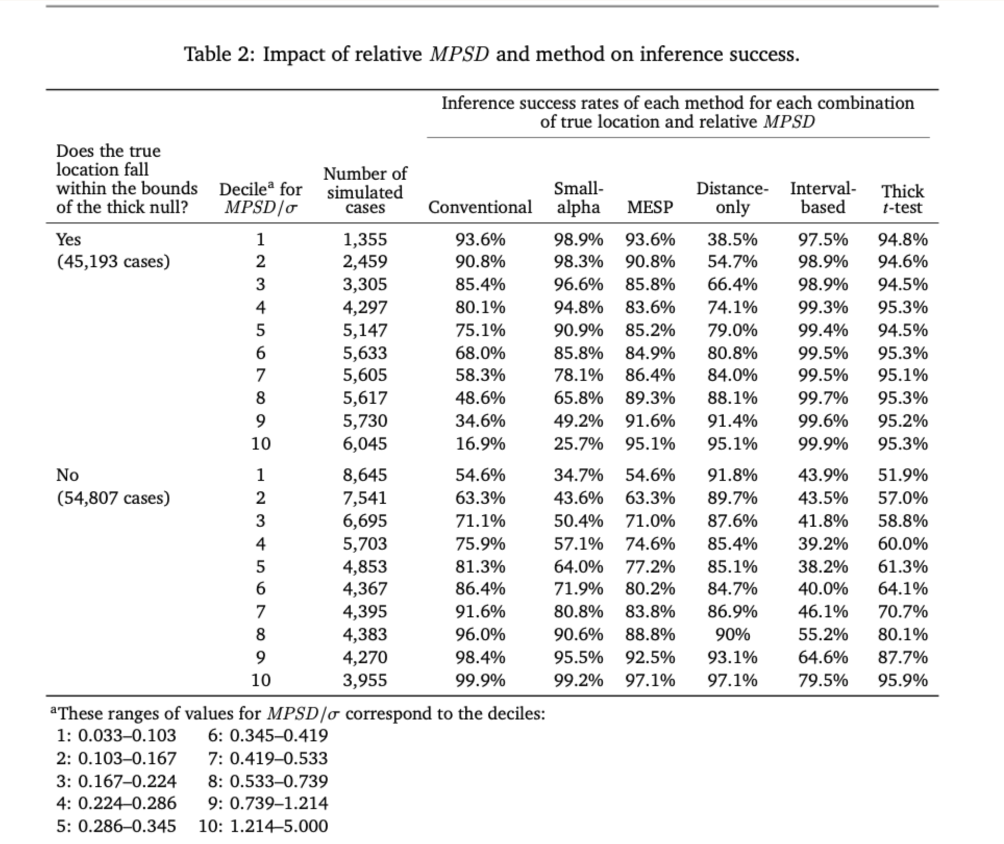


Table 2: Impact of relative *MPSD* and method on inference success.

Does the true location fall within the bounds of the thick null?	Decile[a] for $MPSD/\sigma$	Number of simulated cases	Inference success rates of each method for each combination of true location and relative *MPSD*					
			Conventional	Small-alpha	MESP	Distance-only	Interval-based	Thick t-test
Yes (45,193 cases)	1	1,355	93.6%	98.9%	93.6%	38.5%	97.5%	94.8%
	2	2,459	90.8%	98.3%	90.8%	54.7%	98.9%	94.6%
	3	3,305	85.4%	96.6%	85.8%	66.4%	98.9%	94.5%
	4	4,297	80.1%	94.8%	83.6%	74.1%	99.3%	95.3%
	5	5,147	75.1%	90.9%	85.2%	79.0%	99.4%	94.5%
	6	5,633	68.0%	85.8%	84.9%	80.8%	99.5%	95.3%
	7	5,605	58.3%	78.1%	86.4%	84.0%	99.5%	95.1%
	8	5,617	48.6%	65.8%	89.3%	88.1%	99.7%	95.3%
	9	5,730	34.6%	49.2%	91.6%	91.4%	99.6%	95.2%
	10	6,045	16.9%	25.7%	95.1%	95.1%	99.9%	95.3%
No (54,807 cases)	1	8,645	54.6%	34.7%	54.6%	91.8%	43.9%	51.9%
	2	7,541	63.3%	43.6%	63.3%	89.7%	43.5%	57.0%
	3	6,695	71.1%	50.4%	71.0%	87.6%	41.8%	58.8%
	4	5,703	75.9%	57.1%	74.6%	85.4%	39.2%	60.0%
	5	4,853	81.3%	64.0%	77.2%	85.1%	38.2%	61.3%
	6	4,367	86.4%	71.9%	80.2%	84.7%	40.0%	64.1%
	7	4,395	91.6%	80.8%	83.8%	86.9%	46.1%	70.7%
	8	4,383	96.0%	90.6%	88.8%	90%	55.2%	80.1%
	9	4,270	98.4%	95.5%	92.5%	93.1%	64.6%	87.7%
	10	3,955	99.9%	99.2%	97.1%	97.1%	79.5%	95.9%

[a] These ranges of values for $MPSD/\sigma$ correspond to the deciles:
1: 0.033–0.103
2: 0.103–0.167
3: 0.167–0.224
4: 0.224–0.286
5: 0.286–0.345
6: 0.345–0.419
7: 0.419–0.533
8: 0.533–0.739
9: 0.739–1.214
10: 1.214–5.000

Table 2 presents the impact of relative Minimum Detectable Effect (MDE) and methods on the inference success rates. It categorizes the data based on the true parameter position and the decile of relative MDE. Each method's inference success rate is shown across different scenarios. The table allows for comparison of inference success rates across various relative MDE deciles and true parameter positions, aiding in understanding method performance under different conditions. Overall, it provides valuable insights into the relationship between relative MDE and method effectiveness in inference analysis.

Replication

Implementation Steps:

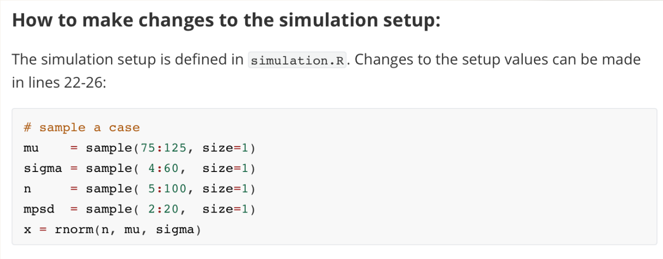
1. Modify simulation.R: Enhance to generate skewness, kurtosis, and interaction terms. Update to accurately reflect distributional characteristics and capture variable interactions.

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.

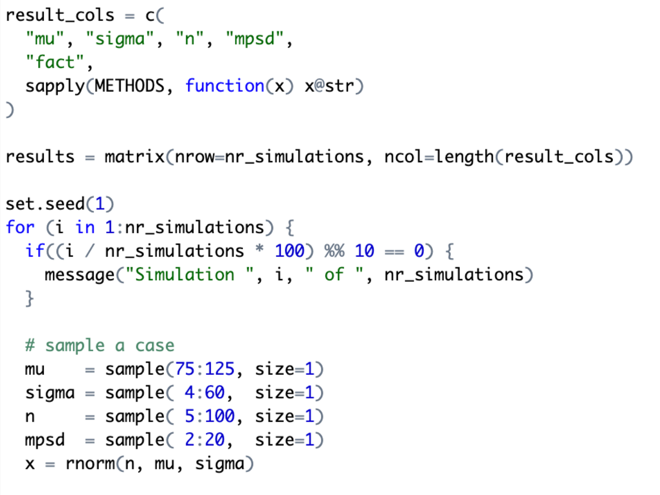
5. Interpretation and Discussion: Analyze results, discuss implications of added covariates, and identify scenarios where criterion performs differently.



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:

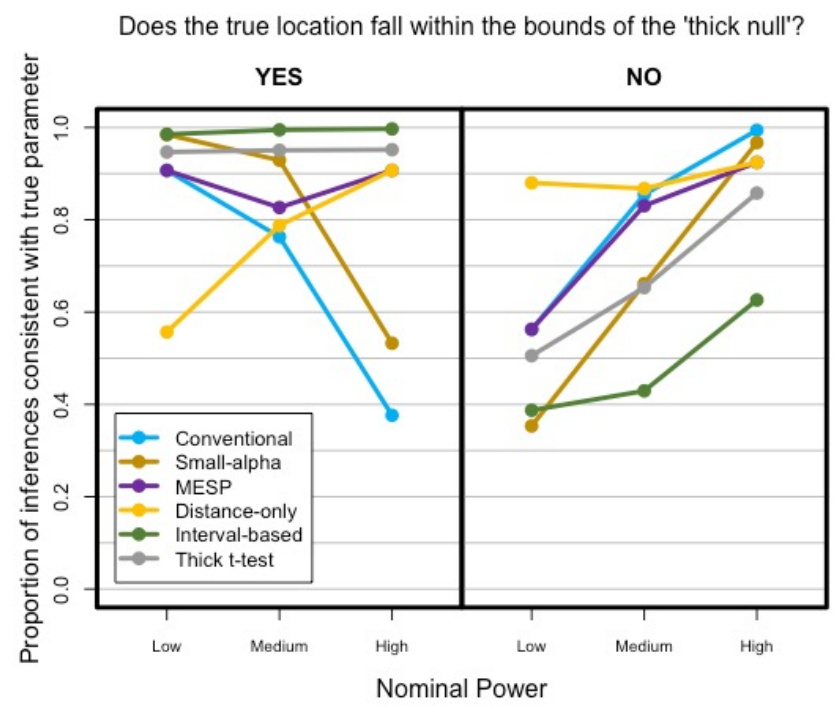
```
# sample a case
mu    = sample(75:125, size=1)
sigma = sample( 4:60,  size=1)
n     = sample( 5:100, size=1)
mpsd  = sample( 2:20,  size=1)
x = rnorm(n, mu, sigma)
```

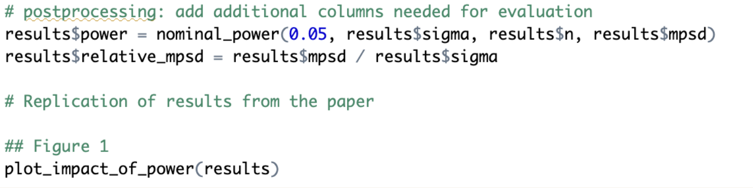


```
result_cols = c(
  "mu", "sigma", "n", "mpsd",
  "fact",
  sapply(METHODS, function(x) x@str)
)

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)
```

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).

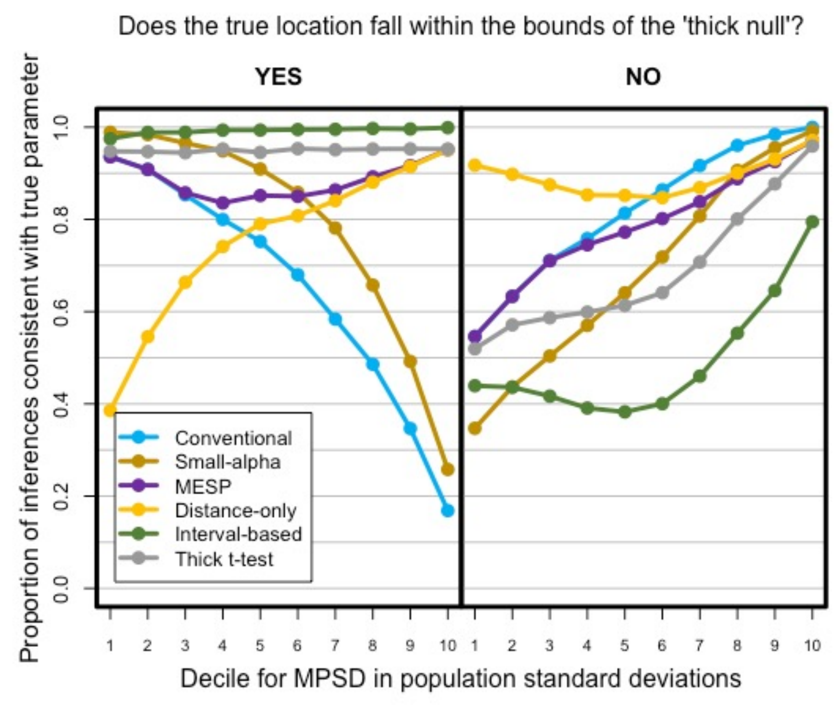
Y-axis: Proportion of inferences consistent with true parameter (indicating the success rate of inference).

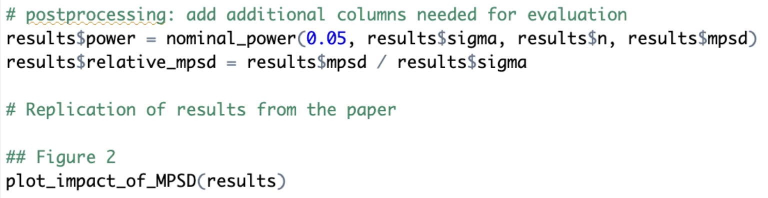
Five Lines:

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)
```

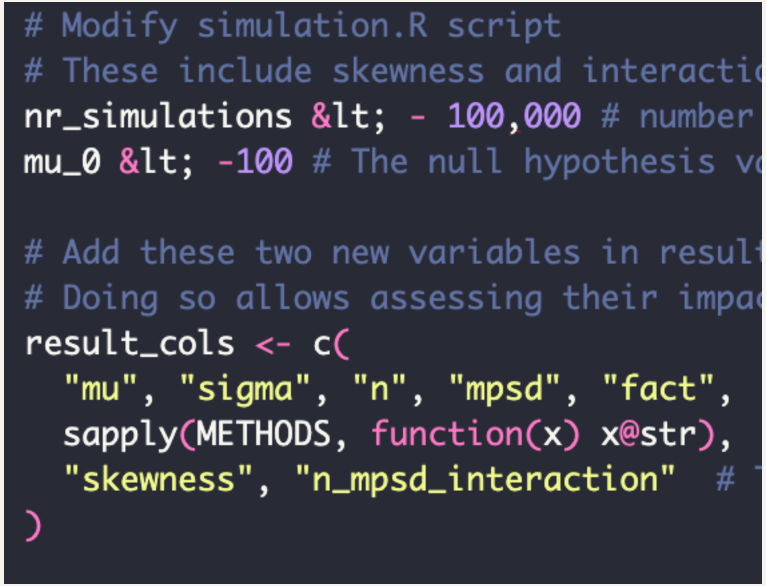
X-axis: Decile for MPSD in population standard deviations. Represents the range of relative Minimum Detectable Effect (MPSD).
Y-axis: Proportion of inferences consistent with true parameter. Indicates the proportion of inference results consistent with the true parameter, i.e., inference success rate.

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

Two additional covariates were introduced:

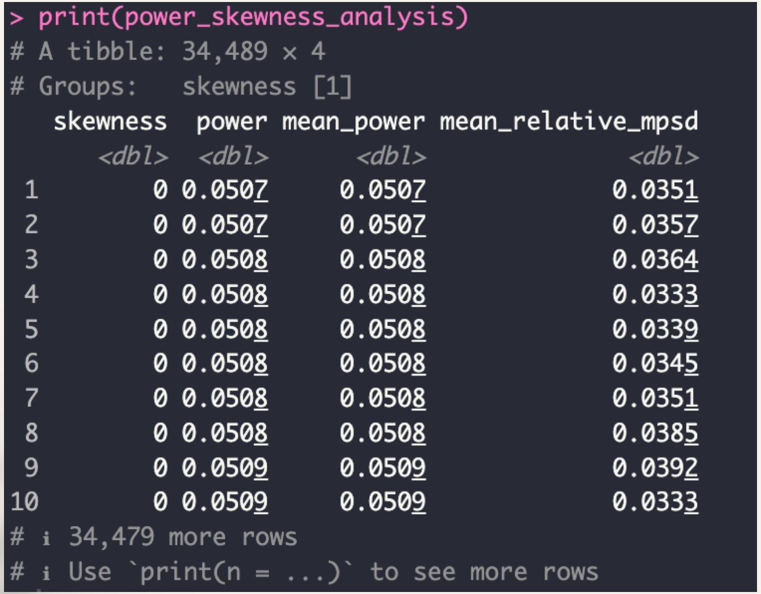


```
# Modify simulation.R script
# These include skewness and interacti
nr_simulations &lt; - 100,000 # number
mu_0 &lt; -100 # The null hypothesis v

# Add these two new variables in resul
# Doing so allows assessing their impa
result_cols <- c(
  "mu", "sigma", "n", "mpsd", "fact",
  sapply(METHODS, function(x) x@str),
  "skewness", "n_mpsd_interaction"  #
)
```

- Skewness: Measures the degree to which the data distribution deviates from symmetry. Introducing skewness allows for the assessment of how skewed data may impact test results.

- Interaction term between sample size and the minimum detectable effect (MDE): Explores the interaction effect between sample size and the minimum detectable effect by their product term.

Impact of Power and Skewness



```
> print(power_skewness_analysis)
# A tibble: 34,489 × 4
# Groups:   skewness [1]
   skewness  power mean_power mean_relative_mpsd
      <dbl>  <dbl>      <dbl>              <dbl>
 1        0 0.0507     0.0507             0.0351
 2        0 0.0507     0.0507             0.0357
 3        0 0.0508     0.0508             0.0364
 4        0 0.0508     0.0508             0.0333
 5        0 0.0508     0.0508             0.0339
 6        0 0.0508     0.0508             0.0345
 7        0 0.0508     0.0508             0.0351
 8        0 0.0508     0.0508             0.0385
 9        0 0.0509     0.0509             0.0392
10        0 0.0509     0.0509             0.0333
# ℹ 34,479 more rows
# ℹ Use `print(n = ...)` to see more rows
```

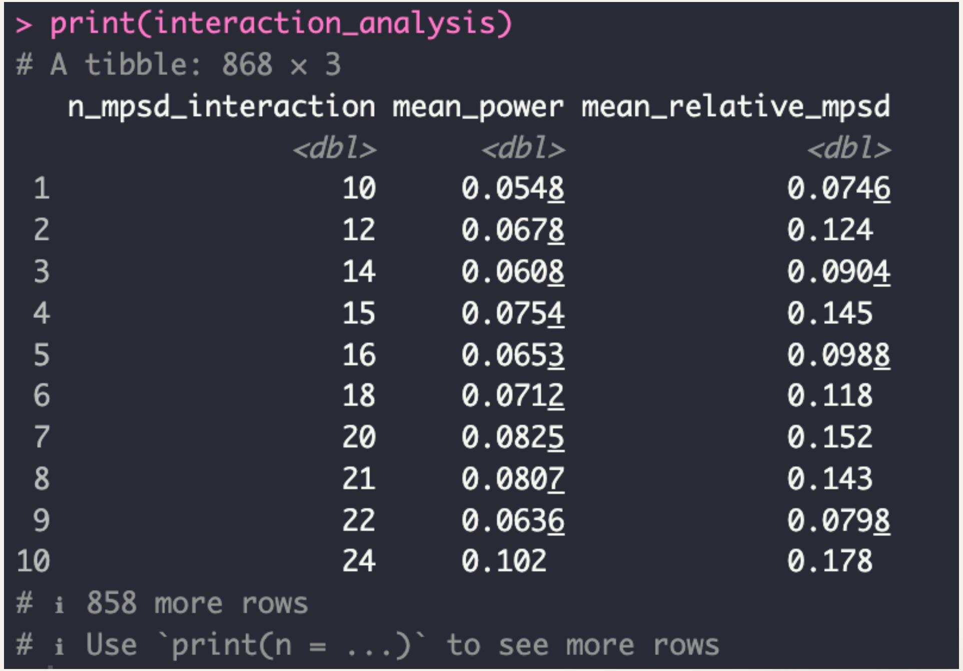
skewness: Represents the skewness value and describes the symmetry of the data distribution.
power column: Represents the power value, i.e. the ability to reject the null hypothesis at a given level of significance.

mean_power: Represents average power, i.e. the average power under a given condition.

mean_relative_mpsd: Indicates the mean relative least significant difference, that is, the mean of the ratio of the least significant difference to the standard deviation.

Conclusion: This reduces the Type II errors (i.e., false negatives) committed when accepting the null hypothesis and improves the accuracy and reliability of the experiment.
Increased reliability of the experiment: High-power experiments have a higher signal-to-noise ratio, so the results are more credible and convincing. This helps to enhance the reproducibility and reliability of the study.

Impact of n*MPSD Interaction:

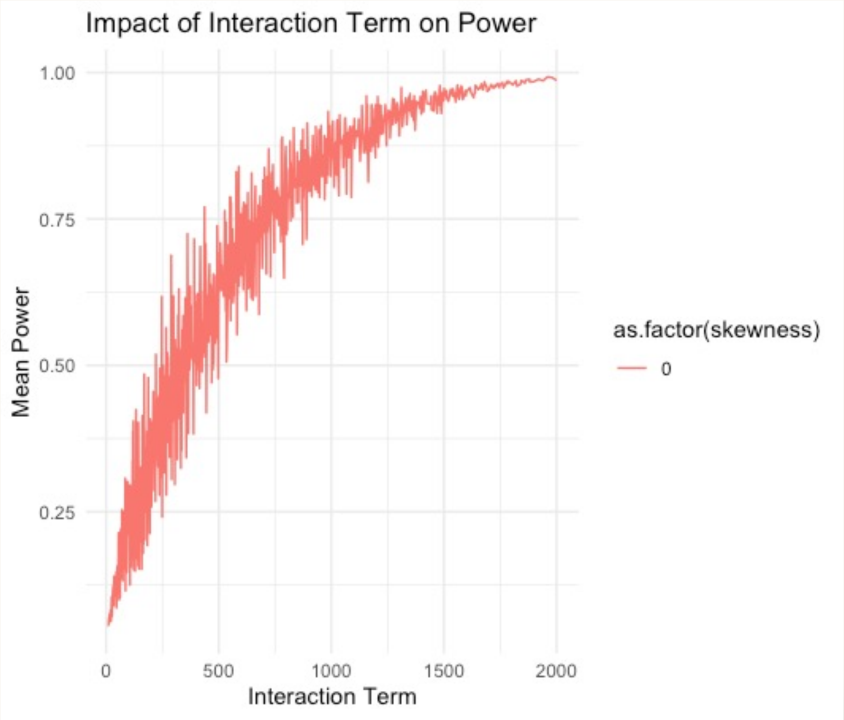


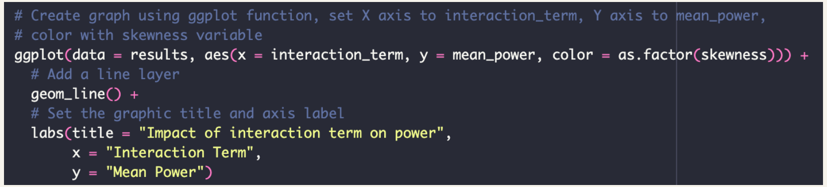
```
> print(interaction_analysis)
# A tibble: 868 × 3
   n_mpsd_interaction mean_power mean_relative_mpsd
                <dbl>      <dbl>              <dbl>
 1                 10     0.0548             0.0746
 2                 12     0.0678             0.124
 3                 14     0.0608             0.0904
 4                 15     0.0754             0.145
 5                 16     0.0653             0.0988
 6                 18     0.0712             0.118
 7                 20     0.0825             0.152
 8                 21     0.0807             0.143
 9                 22     0.0636             0.0798
10                 24     0.102              0.178
# ℹ 858 more rows
# ℹ Use `print(n = ...)` to see more rows
```

About Experimental performance (mean_power) :
With the increase of interaction terms, the average efficiency of the experiment showed a certain trend of change. As the interaction terms increase, the average power may increase, indicating that experimental performance may improve accordingly.

On the perception of minimal significant difference (mean_relative_mpsd) :
Similarly, as the number of interaction terms increases, so does the perception of the least significant difference. An increase in interaction terms may lead to an increase or decrease in the degree of perception of minimal significant differences.

Interact-Power Analysis





```
# Create graph using ggplot function, set X axis to interaction_term, Y axis to mean_power,
# color with skewness variable
ggplot(data = results, aes(x = interaction_term, y = mean_power, color = as.factor(skewness))) +
  # Add a line layer
  geom_line() +
  # Set the graphic title and axis label
  labs(title = "Impact of interaction term on power",
       x = "Interaction Term",
       y = "Mean Power")
```

This diagram shows a linear plot of the effect of the interaction terms on the power. It was observed that with the increase of interaction terms, the average power showed a trend of first sharp increase, and then gradually slow down, and finally stabilized at 1. The X-axis represents the value of the interaction item, ranging from 0 to 2000. The Y-axis is the value of the average power, ranging from 0 to 1. The entire graph is "the effect of the interaction term on the power", and the lines in the graph represent changes in different skewness levels.

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