Combining Parametric and Non-Parametric Effect Sizes to Create a Powerful Two-Sample Test

Nick Chandler, Fiona Cleary, Sam Burnett, Kimihiro Noguchi*

Department of Mathematics, Western Washington University

*noguchk@wwu.edu

July 11, 2025

Overview

Motivating Example

Parametric and Non-Parametric Tests

Location-Scale Families

Our Research

Motivating Example

Motivating Example

Consider the lengths of remission for leukemia patients under two different control and test drugs (Maurya et al., 2011).

- Number of treatments: 4 (2 control drugs and 2 test drugs)
- Number of patients per treatment: 20

We could perform 6 all-pairwise comparisons of these control and test drugs utilizing both the t-test and the Brunner-Munzel (BM) test (Brunner and Munzel, 2000).

Histogram

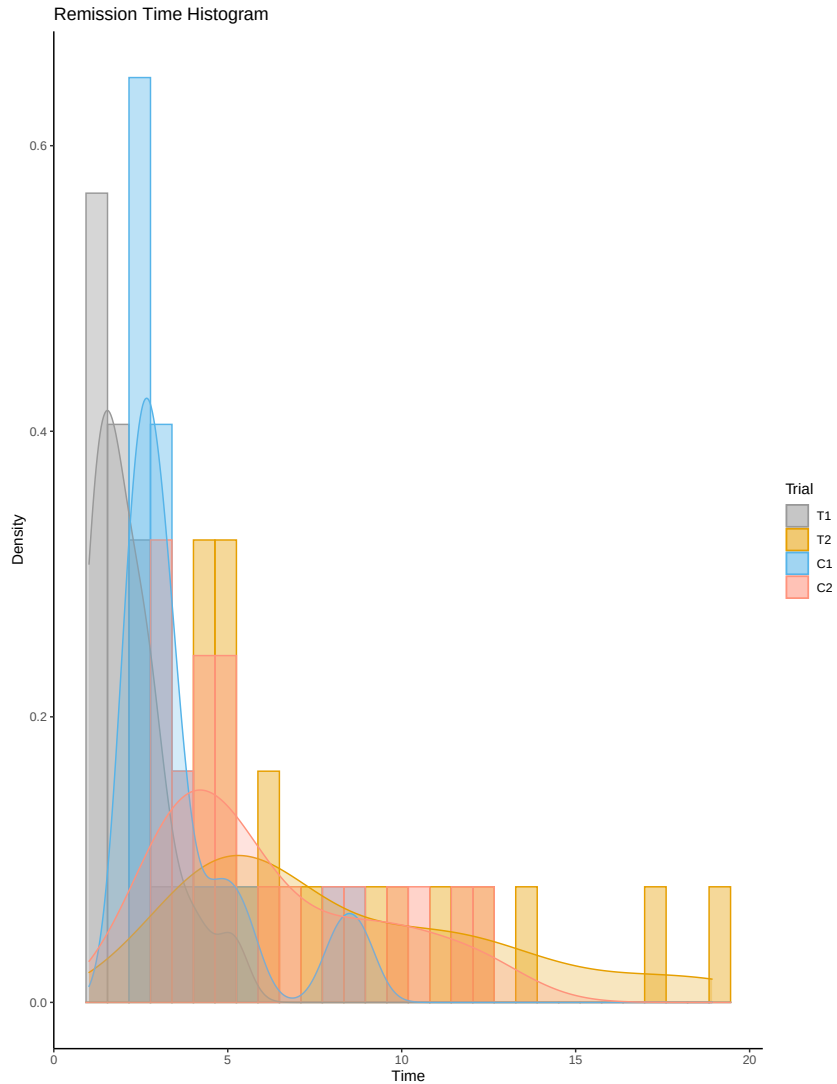


Figure: Histograms of the two control drugs (C1 and C2) and two test drugs (T1 and T2), showing skewness with some possible location shift.

Leukemia Data

Statistic	Test Drug 1 (T1)	Test Drug 2 (T2)	Control Drug 1 (C1)	Control Drug 2 (C2)
Sample Size	20	20	20	20
Sample Mean	2.189	8.369	3.668	6.143
Sample Minimum	1.013	4.498	2.214	3.071
Sample Scale	1.176	3.871	1.454	3.072

Table: Summary of important statistics including the sample size, mean, minimum, and scale for each treatment based on Maurya et al. (2011).

Observations:

- Sample Mean: $T2 > C2 >> C1 > T1$
- Sample Scale: $T2 > C2 >> C1 > T1$

Parametric and Non-Parametric Tests

Parametric Effect Size: Mean Difference

Let $\mathbf{X} = \{X_1, \ldots, X_{n_X}\} \sim F_X$ and $\mathbf{Y} = \{Y_1, \ldots, Y_{n_Y}\} \sim F_Y$ be independent random samples from some continuous distributions F_X and F_Y with $\mu_X = E[X_i]$, $\mu_Y = E[Y_j]$, $\sigma_X^2 = \text{Var}(X_i)$, and $\sigma_Y^2 = \text{Var}(Y_j)$.

Furthermore, let $\overline{X}$, $\overline{Y}$, S_X^2, and S_Y^2 be the sample means and variances of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Parametric Effect Size (Mean Difference): $\mu_d = \mu_X - \mu_Y$.

Sample Mean Difference: $\hat{\mu}_d = \overline{X} - \overline{Y}$.

Common Parametric Test: Welch's T-Test

To test $H_0 : \mu_X - \mu_Y = 0$, we use Welch's test (heteroscedastic t-test):

$$T = \frac{\hat{\mu}_d - \mu_d}{\sqrt{S_X^2/n_X + S_Y^2/n_Y}}.$$

The distribution of T under H_0 can be approximated by the t-distribution with ν_T degrees of freedom, where

$$\nu_T = \frac{(S_X^2/n_X + S_Y^2/n_Y)^2}{S_X^4/[n_X^2(n_X - 1)] + S_Y^4/[n_Y^2(n_Y - 1)]}.$$

Non-Parametric Effect Size: Relative Effect

Relative Effect/Stochastic Superiority:

$$d = \Pr(X_i < Y_j) + 0.5 \Pr(X_i = Y_j).$$

For continuous distributions, $d = \Pr(X_i < Y_j)$ as $\Pr(X_i = Y_j) = 0$.

Interpretation:

- $d < 0.5$: F_X is stochastically superior to F_Y.
- $d = 0.5$: F_X and F_Y are stochastically equal.
- $d > 0.5$: F_X is stochastically inferior to F_Y.

Estimating Relative Effect

First, combine $\mathbf{X}$ and $\mathbf{Y}$, and convert it into ranks $\mathbf{R}_X$ and $\mathbf{R}_Y$.

Example: $\mathbf{X} = (0, 3, 2)$ and $\mathbf{Y} = (1, -5)$. Then, $(\mathbf{X}, \mathbf{Y}) = (0, 3, 2, 1, -5)$ and the rank transformation gives $(2, 5, 4, 3, 1)$ so that $\mathbf{R}_X = (2, 5, 4)$ and $\mathbf{R}_Y = (3, 1)$.

Let $\overline{R}_X$ and $\overline{R}_Y$ be the sample mean of the ranks from $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Sample Relative Effect:

$$\hat{d} = \frac{1}{N}(\overline{R}_Y - \overline{R}_X) + 0.5,$$

where $N = n_X + n_Y$.

Common Non-Parametric Test: Brunner-Munzel Test

To test $H_0 : d = 0.5$, we use the Brunner-Munzel test (heteroscedastic Wilcoxon-Mann-Whitney test).

$$W = \frac{\hat{d} - d}{\sqrt{S_{R_X}^2/n_X + S_{R_Y}^2/n_Y}},$$

where $S_{R_X}^2$ and $S_{R_Y}^2$ are the sample variances of $\mathrm{Var}(F_Y(X_i))$ and $\mathrm{Var}(F_X(Y_j))$.

The distribution of W under H_0 can be approximated by the t-distribution with ν_W degrees of freedom, where

$$\nu_W = \frac{(S_{R_X}^2/n_Y + S_{R_Y}^2/n_X)^2}{S_{R_X}^4/[n_Y^2(n_Y - 1)] + S_{R_Y}^4/[n_X^2(n_X - 1)]}.$$

Location-Scale Families

Location-Scale Families

Definition: Let $X \sim F_X$. Then, F_X is said to belong to location-scale families if, for $Z = (X - \mu_X)/\sigma_X$,

$$F_X^{-1}(p) = \mu_X + \sigma_X F_Z^{-1}(p),$$

$p \in (0, 1)$, $\mu_X \in \mathbb{R}$, $\sigma_X > 0$.

Example #1: Normal Distribution Family $N(\mu_X, \sigma_X^2)$.

Example #2: Two-Parameter Exponential Distribution Family $\text{Exp}(L_X, \theta_X)$, where $L_X \in \mathbb{R}$ denotes the lower bound and $\theta_X > 0$ denotes the scale parameter.

Assumptions

Let $X \sim F_X$ and $Y \sim F_Y$, where F_X and F_Y are assumed to belong to the same location-scale family. Then,

$$F_X^{-1}(b) = \mu_X + \sigma_X F_Z^{-1}(b) \text{ and } F_Y^{-1}(b) = \mu_Y + \sigma_Y F_Z^{-1}(b)$$

for all $b \in (0, 1)$, $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma_X, \sigma_Y > 0$.

Our Research

The Test

The test we propose is one which combines the two effect size measures discussed previously into one test. This seeks to quell the controversy over which test to use.

- Combination of the two-sample t-test and Brunner-Munzel test
- Detects a significant mean difference and/or deviation from the stochastic equality
- Allows for an easier and powerful comparison between two groups of a study

The Test

Let $\hat{\mu}$, $\hat{\sigma}$, and $\hat{\sigma}_X$ denote the maximum likelihood estimate (MLE) of $\mu = (\mu_Y - \mu_X)/\sigma_X$, $\sigma = \sigma_Y/\sigma_X$, and σ_X, respectively. Further, let $D_b(\mu, \sigma)$ and $K_d(\mu, \sigma)$ denote functions which measure the mean difference and stochastic inequality, respectively. Then, the null hypothesis of our test is:

$$H_0 \colon \{\mu_d = 0\} \cap \{d = 0.5\},$$

We can test this using the joint test statistics,

$$(T_1, T_2)' = \left(\frac{\hat{\sigma}_X D_b(\hat{\mu}, \hat{\sigma})}{\sqrt{\hat{\sigma}_{11}/N}}, \frac{\hat{\sigma}_X K_d(\hat{\mu}, \hat{\sigma})}{\sqrt{\hat{\sigma}_{22}}/N} \right)',$$

which are asymptotically bivariate normal under H_0.

Our Research

Our research is ongoing.

- There is mathematical justification for this test (Noguchi et al., 2023)
- Currently, we are evaluating its characteristics relative to the t-test and Brunner-Munzel test
- This is done using a High-Throughput Computing environment to run simulations

The Study

The simulation study we are conducting has a few main objectives, focused on measuring the out-performance of the traditional tests discussed earlier:

Robustness:

- The probability of falsely detecting a mean difference and/or stochastic inequality when there is none.
- This is known as a Type I error rate or a false positive rate.

Power:

- The probability of correctly detecting a mean difference and/or stochastic inequality when there is one.

The Study

Methods:

- We compare a variety of bootstrap methods to measure the statistical significance and determine the best version of the test.
- We utilize Western Washington University's College of Science of Engineering computing cluster to run many jobs in parallel.

Goals:

- Where/how does our test perform better than the two-sample t-test?
- The Brunner-Munzel test?

Bootstrap Overview

Bootstrapping is a resampling technique for computing p-values in inferential statistics:

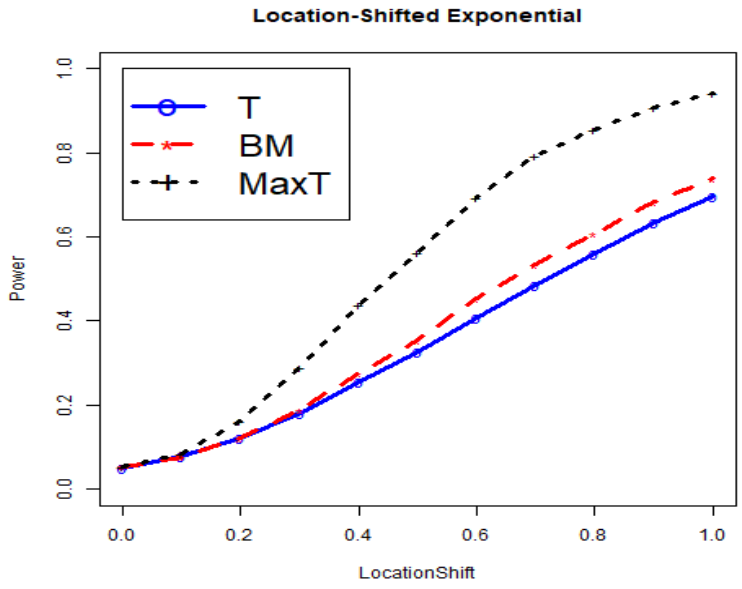
- A resample is created by taking observations from the sample which respects the null hypothesis with replacement. Each resample has the same sample size as the original sample.
- Many resamples are created and the test statistics are computed from each resample.
- These test statistics from the resamples approximate the distribution of the test statistic under the null hypothesis.
- The p-value is computed by calculating the proportion of these bootstrapped test statistics exceeding the original test statistic.

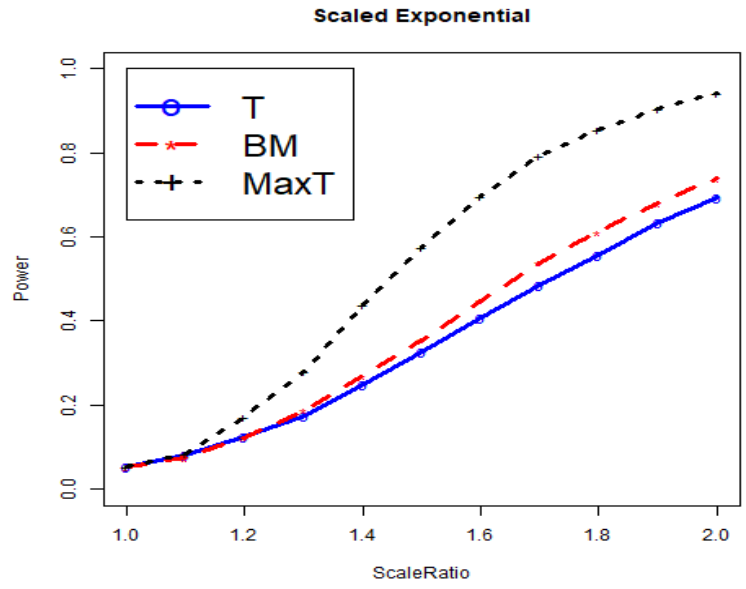
Max-T Method

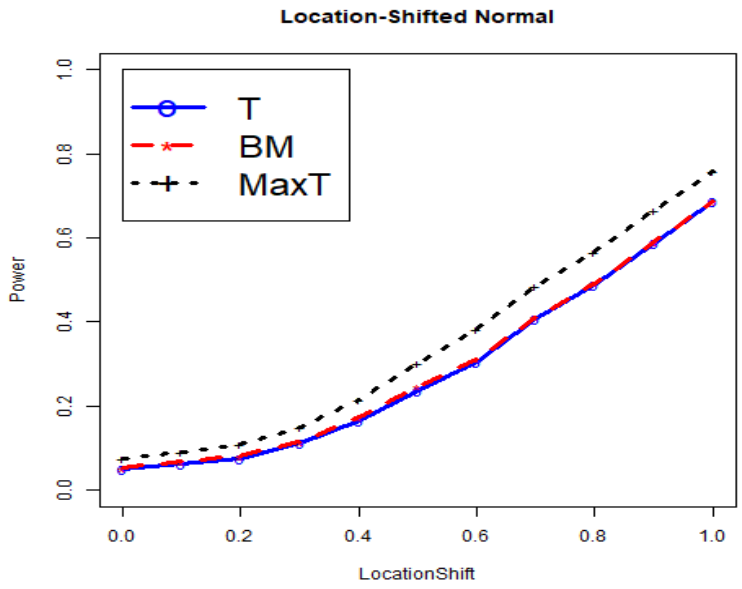
Overview

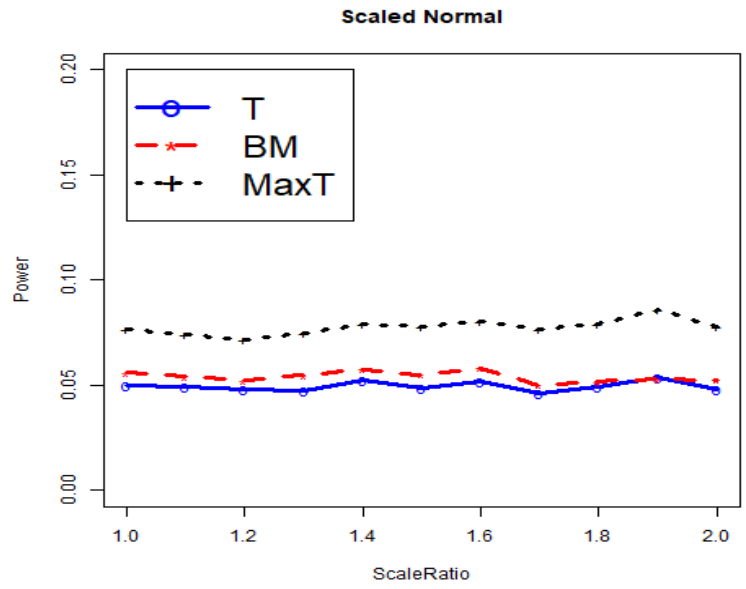
- Max-T looks at the maximum of n test statistics ($n = 2$ in our case, namely, $|T_1|$ and $|T_2|$).
  That is, we compute $T_{\max} = \max\{|T_1|, |T_2|\}$.
- Bootstrapped $T_{\max}$ statistics are computed and the proportion of bootstrapped $T_{\max}$ statistics exceeding the $T_{\max}$ statistic from the original data is the computed p-value.

Power Curve Plots









Simulation Results

Our simulation study shows some promising results:

- The robustness of our test is reasonable at $\alpha = 0.05$ for the two-parameter exponential distribution.
- On the other hand, our test tends to be liberal (> 0.05) for the normal distribution, which needs to be addressed in the future.
- Our test tends to be way more powerful than the two-sample t-test and the Brunner-Munzel test for the two-parameter exponential distribution.

Further Remarks on the Simulation Study

A few comments on the future work:

- Different bootstrap methods need be explored to ensure that our test is robust under the normal distribution.
- Note that our simulation set-up is for detecting a location shift and/or a scale difference. However, our test is developed for detecting a mean difference and/or relative effect (stochastic superiority). Thus, a more relevant simulation design can help us better understand the performance of our test.
- Nevertheless, our test is a close analog to location-scale tests. For our future work, we can compare our test to the location-scale Cucconi test (Marozzi, 2009).

Thank You!

References

1. Brunner, E. and Munzel, U. (2000). The nonparametric Behrens-Fisher Problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal* 42:1, 17–25.
2. Marozzi, M. (2009). Some notes on the location-scale Cucconi test. *Journal of Nonparametric Statistics* 21:5, 629–647.
3. Maurya, V., Goyal, A., and Nath Gill, A. (2011). Multiple comparisons with more than one control for exponential location parameters under heteroscedasticity. *Communications in Statistics – Simulation and Computation* 40:5, 621–644.
4. Noguchi, K., Burnett, S., and Cleary, F. (2023). On the duality between stochastic superiority and equality of quantiles in nonparametric inference. Working paper.