Pragmastat: Pragmatic Statistical Toolkit

Andrey Akinshin
andrey.akinshin@gmail.com

Version 1.0.0

Abstract

This manual presents a unified statistical toolkit for reliable analysis of real-world data. The toolkit nearly matches the efficiency of traditional statistical estimators under normality, has practically reasonable robustness, enables simple software implementations without advanced statistical libraries, and provides clear explanations accessible to practitioners without deep statistical training. The toolkit consists of renamed, recombined, and refined versions of existing methods.

Contents

1 Toolkit **2**

1.1 One-Sample Summary 2

1.2 Two-Sample Summary 3

2 Estimators **3**

2.1 Center 3

2.2 Spread 4

2.3 Volatility 5

2.4 Precision 6

2.5 MedShift 7

2.6 MedRatio 8

2.7 MedSpread 8

2.8 MedDisparity 9

3 Studies **9**

3.1 Asymptotic Gaussian Expected Value of the Spread 10

3.2 Asymptotic Gaussian Efficiency of the Center 10

3.3 Asymptotic Gaussian Efficiency of the Median 12

3.4 Asymptotic Gaussian Efficiency of the Spread 13

3.5 Asymptotic Gaussian Efficiency of the Median Absolute Deviation 14

3.6 Finite-Sample Efficiency of Central Tendency Estimators 15

3.7 Finite-Sample Efficiency of Dispersion Estimators 17

1 Toolkit

This section provides statistical estimators for summarizing a single sample and comparing two samples.

1.1 One-Sample Summary

Consider a sample $\mathbf{x}$ of n real numbers: $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. The toolkit provides four estimators to summarize key properties of the data and provide insights into the data's primary characteristics:

$$\operatorname{Center}(\mathbf{x}) = \operatorname*{Median}_{1 \le i \le j \le n} \left(\frac{x_i + x_j}{2} \right)$$

$$\operatorname{Spread}(\mathbf{x}) = \operatorname*{Median}_{1 \le i < j \le n} |x_i - x_j|$$

$$\operatorname{Volatility}(\mathbf{x}) = \frac{\operatorname{Spread}(\mathbf{x})}{|\operatorname{Center}(\mathbf{x})|}$$

$$\operatorname{Precision}(\mathbf{x}) = \frac{2 \cdot \operatorname{Spread}(\mathbf{x})}{\sqrt{n}}$$

One-sample summary statistics work best for unimodal distributions and distributions with low dispersion.

$\operatorname{Center}(\mathbf{x})$[1] estimates the central (average) value of the distribution. For normal distributions, it matches both the mean and the median. It outperforms traditional estimators in practical use. Compared to Mean, Center is much more robust (tolerates almost one-third of outliers). Compared to Median, Center is much more efficient and requires 1.5 times fewer observations to achieve the same precision.

$\operatorname{Spread}(\mathbf{x})$[2] estimates distribution dispersion (variability or scatter). It measures the median absolute difference between two random sample elements. This measure offers a practical alternative to standard deviation (StdDev) and median absolute deviation (MAD). Compared to StdDev, Spread is more robust (standard deviation breaks with a single extreme value) and has comparable efficiency under normality. Compared to MAD, Spread is much more efficient under normality and requires 2.35 times fewer observations to achieve the same precision.

$\operatorname{Volatility}(\mathbf{x})$ estimates the relative dispersion of the distribution. Convenient to express in percentage: e.g., a value of 0.2 means 20% relative to $\operatorname{Center}(\mathbf{x})$. Volatility is scale-invariant, which makes an experiment design more portable.

$\operatorname{Precision}(\mathbf{x})$ estimates the distance between two Center estimations of independent random samples. The interval $\operatorname{Center}(\mathbf{x}) \pm \operatorname{Precision}(\mathbf{x})$ forms a range that contains the true center value with high confidence. For even higher confidence, use $\operatorname{Center}(\mathbf{x}) \pm 2 \cdot \operatorname{Precision}(\mathbf{x})$ or $\operatorname{Center}(\mathbf{x}) \pm 3 \cdot \operatorname{Precision}(\mathbf{x})$.

These estimators build on $\operatorname{Median}(\mathbf{x})$[3]. To find the median, $\mathbf{x}$ must first be arranged into a sorted sample[4]: $(x_{(1)}, \ldots, x_{(n)})$. In this ordered sequence, $x_{(1)}$ represents the smallest value and $x_{(n)}$ the largest. The median then becomes the middle value of this sorted sample. When the sample size is even, the median equals the average of the two middle values:

$$\operatorname{Median}(\mathbf{x}) = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

[1] Also known as the *Hodges–Lehmann* location estimator, see [HL63], [Sen63]

[2] Also known as the *Shamos* scale estimator, see [Sha76]

[3] Also known as *sample median*

[4] Also known as *order statistics*

1.2 Two-Sample Summary

Consider a second sample $\mathbf{y}$ of m real numbers: $\mathbf{y} = (y_1, \ldots, y_m)$. Estimators to compare $\mathbf{x}$ and $\mathbf{y}$:

$$\operatorname{MedShift}(\mathbf{x}, \mathbf{y}) = \underset{1 \le i \le n,\ 1 \le j \le m}{\operatorname{Median}} (x_i - y_j)$$

$$\operatorname{MedRatio}(\mathbf{x}, \mathbf{y}) = \underset{1 \le i \le n,\ 1 \le j \le m}{\operatorname{Median}} \left(\frac{x_i}{y_j} \right)$$

$$\operatorname{MedSpread}(\mathbf{x}, \mathbf{y}) = \frac{n \operatorname{Spread}(\mathbf{x}) + m \operatorname{Spread}(\mathbf{y})}{n + m}$$

$$\operatorname{MedDisparity}(\mathbf{x}, \mathbf{y}) = \frac{\operatorname{MedShift}(\mathbf{x}, \mathbf{y})}{\operatorname{MedSpread}(\mathbf{x}, \mathbf{y})}$$

These estimators work best for unimodal or narrow distributions, capturing the typical differences between $\mathbf{x}$ and $\mathbf{y}$.

$\operatorname{MedShift}(\mathbf{x}, \mathbf{y})$[5] estimates the median absolute difference between elements of $\mathbf{x}$ and $\mathbf{y}$. It answers "by how much does $\mathbf{x}$ typically exceed $\mathbf{y}$?" in the original units. The sign matters: positive means $\mathbf{x}$ is typically larger, negative means $\mathbf{y}$ is typically larger. E.g., MedShift of -5 means that in 50% of (x_i, y_j) pairs, $y_j - x_i > 5$.

$\operatorname{MedRatio}(\mathbf{x}, \mathbf{y})$[6] estimates the median ratio of $\mathbf{x}$ elements to $\mathbf{y}$ elements. It answers "what's the typical ratio between $\mathbf{x}$ and $\mathbf{y}$?" as a multiplier. For example, MedRatio $= 1.2$ means that in 50% of (x_i, y_j) pairs, x_i is larger than y_j by at least 20%. Express as percentage change: $(\text{MedRatio} - 1) \times 100\%$. MedRatio is scale-invariant, which makes an experiment design more portable.

$\operatorname{MedSpread}(\mathbf{x}, \mathbf{y})$ estimates the averaged variability when considering both samples together. The measure computes the weighted average of individual spreads, where larger samples contribute more. This value primarily serves as a scaling factor for MedDisparity. It represents the typical variability in the combined data and works best for distributions with similar dispersion values.

$\operatorname{MedDisparity}(\mathbf{x}, \mathbf{y})$[7] estimates a normalized absolute difference between $\mathbf{x}$ and $\mathbf{y}$ expressed in standardized spread units. Negative values are treated similarly to $\operatorname{MedShift}(\mathbf{x}, \mathbf{y})$. MedDisparity is scale-invariant, which makes an experiment design more portable.

2 Estimators

This section explains each estimator's key properties, selection rationale, and advantages over traditional methods.

2.1 Center

$$\operatorname{Center}(\mathbf{x}) = \underset{1 \le i \le j \le n}{\operatorname{Median}} \left(\frac{x_i + x_j}{2} \right)$$

Practical Recommendations

Center provides an initial insight into the magnitude of sample values. When Volatility is small, the whole sample can be approximated by the Center value.

Key Facts

[5] Also known as the *Hodges–Lehmann shift estimator*

[6] Inspired by the *Hodges–Lehmann estimator*

[7] A robust alternative to traditional effect size measures like Cohen's d

- Measures central tendency (the average value)
- Domain: any real numbers
- Equals the *Hodges-Lehmann estimator* ([HL63], [Sen63]), renamed to "Center" for clarity
- Called *pseudomedian* in some texts because it is consistent with the median for symmetric distributions
- Asymptotic Gaussian efficiency: $\approx 96\%$; finite-sample Gaussian efficiency $> 91\%$
- Asymptotic breakdown point: $\approx 29\%$

Comparison

- Compared to the *mean*: more robust (tolerates almost one-third of outliers) and has comparable efficiency under normality
- Compared to the *median*: more efficient under normality and requires 1.5 times fewer observations for the same precision

Properties

$$\text{Center}(\mathbf{x} + k) = \text{Center}(\mathbf{x}) + k$$

$$\text{Center}(k \cdot \mathbf{x}) = k \cdot \text{Center}(\mathbf{x})$$

2.2 Spread

$$\text{Spread}(\mathbf{x}) = \underset{i<j}{\text{Median}} |x_i - x_j|$$

Practical Recommendations

Spread provides an initial insight into the dispersion of the sample values. Interpretation: half of $|x_i - x_j|$ is smaller than $\text{Spread}(\mathbf{x})$, the other half is larger.

Key Facts

- Measures dispersion (also known as variability or scatter)
- Domain: any real numbers (for $n = 1$, it is convenient to use $\text{Spread}(\mathbf{x}) = 0$)
- Equals the *Shamos estimator* ([Sha76]), renamed to "Spread" for clarity
- Asymptotic Gaussian efficiency: $\approx 86\%$
- Asymptotic breakdown point: $\approx 29\%$ (matches Center in robustness)
- Asymptotic expected value for the standard normal distribution: ≈ 0.954
- Not consistent for the standard deviation under normality

Comparison

- Compared to the *standard deviation*: more robust (tolerates almost one-third of outliers) and has comparable efficiency under normality; more intuitive without requiring knowledge of normal distributions
- Compared to the *median absolute deviation*: more efficient under normality and requires ≈ 2.35 times fewer observations for the same precision

Empirical Rule

For the standard normal distribution, the asymptotic 68–95–99.7 rule becomes the 66-94-99.6 rule:

- $[\text{Center}(\mathbf{x}) \pm 1 \cdot \text{Spread}(\mathbf{x})]$ covers $\approx 65.98518\%$ of the distribution

- $[\text{Center}(\mathbf{x}) \pm 2 \cdot \text{Spread}(\mathbf{x})]$ covers $\approx 94.35758\%$ of the distribution
- $[\text{Center}(\mathbf{x}) \pm 3 \cdot \text{Spread}(\mathbf{x})]$ covers $\approx 99.57851\%$ of the distribution
- $[\text{Center}(\mathbf{x}) \pm 4 \cdot \text{Spread}(\mathbf{x})]$ covers $\approx 99.98641\%$ of the distribution
- $[\text{Center}(\mathbf{x}) \pm 5 \cdot \text{Spread}(\mathbf{x})]$ covers $\approx 99.99982\%$ of the distribution

Properties

$$\text{Spread}(\mathbf{x} + k) = \text{Spread}(\mathbf{x})$$

$$\text{Spread}(k \cdot \mathbf{x}) = |k| \cdot \text{Spread}(\mathbf{x})$$

$$\text{Spread}(x) \geq 0$$

2.3 Volatility

$$\text{Volatility}(\mathbf{x}) = \frac{\text{Spread}(\mathbf{x})}{|\text{Center}(\mathbf{x})|}$$

Practical Recommendations

Volatility provides a scale-invariant insight into the distribution dispersion normalized by the center value.

Interpretation examples: - $\text{Volatility}(\mathbf{x}) = 1\%$: data clusters tightly around $\text{Center}(\mathbf{x})$ with minimal variation - $\text{Volatility}(\mathbf{x}) = 10\%$: moderate variation, typical values range from 90% to 110% of center - $\text{Volatility}(\mathbf{x}) = 100\%$: high variation, values span from near zero to twice the center

Key Facts

- Measures the relative dispersion of a sample to $\text{Center}(\mathbf{x})$
- Domain:
  - Mathematical Domain: $Center(\mathbf{x}) \neq 0$
  - Logical Domain: all sample elements have the same sign, sample doesn't contain zeros
  - Pragmatic Domain: non-negative values allowing up to 29% zeros
- Robust alternative to the *coefficient of variation*
- Scale-invariant, which makes an experiment design more portable

Properties

$$\text{Volatility}(k \cdot \mathbf{x}) = \text{Volatility}(\mathbf{x})$$

$$\text{Volatility}(x) \geq 0$$

2.4 Precision

$$\operatorname{Precision}(\mathbf{x}) = \frac{2 \cdot \operatorname{Spread}(\mathbf{x})}{\sqrt{n}}$$

Practical Recommendations

The interval $\operatorname{Center}(\mathbf{x}) \pm k \cdot \operatorname{Precision}(\mathbf{x})$ contains the true center value with probability depending on k. Select $k \in \{1, 2, 3\}$ based on required confidence.

Low error costs: $\operatorname{Center}(\mathbf{x}) \pm \operatorname{Precision}(\mathbf{x})$ provides a reasonable interval.
High error costs: repeat the experiment 3–5 times and use $\operatorname{Center}(\mathbf{x}) \pm 2 \cdot \operatorname{Precision}(\mathbf{x})$.
Critical applications: repeat the experiment 7–10 times and use $\operatorname{Center}(\mathbf{x}) \pm 3 \cdot \operatorname{Precision}(\mathbf{x})$.

Key Facts

- Measures how much $\operatorname{Center}(\mathbf{x})$ would fluctuate from sample to sample if samples of the same size were repeatedly drawn under the same conditions
- Domain: any real numbers
- Can be perceived as half of a confidence interval with high confidence level

Properties

$$\operatorname{Precision}(\mathbf{x} + k) = \operatorname{Precision}(\mathbf{x})$$

$$\operatorname{Precision}(k \cdot \mathbf{x}) = |k| \cdot \operatorname{Precision}(\mathbf{x})$$

$$\operatorname{Precision}(\mathbf{x}) \geq 0$$

Comments on default factor of 2

Practitioners tend to perceive estimations $a \pm b$ as “the value is inside $[a - b; a + b]$” ignoring uncertainty. The unscaled interval $\operatorname{Center}(x) \pm \operatorname{Spread}(\mathbf{x})/\sqrt{n}$ covers the true distribution center value only in $\approx 65\%$ of cases. To reduce risks of misinterpretation of $\operatorname{Center}(\mathbf{x}) \pm \operatorname{Precision}(\mathbf{x})$, it is reasonable to use the default factor for $\operatorname{Precision}(\mathbf{x})$ to ensure high coverage of such intervals for the true distribution center value. For simpler calculations, it’s convenient to use natural numbers as scale factors for $\operatorname{Precision}(\mathbf{x})$. The factor of 2 is chosen since it’s the smallest natural number that ensures decent coverage. Natural coefficients k produce a standardized discrete precision scale with values of $k \cdot \operatorname{Precision}(\mathbf{x})$ or $2k \cdot \operatorname{Spread}(\mathbf{x})/\sqrt{n}$.

Relation between Precision and Confidence Intervals

In the strict normal model, it’s convenient to express precision via the *standard error* (the standard deviation divided by the square root of sample size). The standard error can be scaled to the margin of error (half of a confidence interval) for the given confidence level. Choosing between confidence levels — 95%, 99%, 99.9%, or even 89%[8] — remains arbitrary. Practitioners struggle to extract insights from confidence intervals and levels without calculation tools. This difficulty leads to frequent misinterpretation because no standard exists for choosing levels. Knowing one confidence interval determines all others through constant relationships, yet practitioners report two numbers (interval size and confidence level) that are hard to comprehend together. A single standardized value would serve practitioners better.

A standardized value simplifies reporting and improves consistency. A memorable definition enables mental calculation without advanced tools. Practitioners develop intuition through repeated use.

[8] See https://github.com/easystats/bayestestR/discussions/250

Precision has no direct mapping into traditional confidence intervals. Confidence intervals ensure the declared coverage rate only under perfect normality. Real data provides no guarantees that a 99% confidence interval actually covers the true value in 99% of experiments. The table below shows the translation from $\mathrm{Center}(\mathbf{x}) \pm k \cdot \mathrm{Precision}(\mathbf{x})$ to confidence levels under normality:

n	k=1	k=2	k=3
2	0.78364	0.88862	0.92534
3	0.88660	0.96741	0.98507
4	0.89523	0.98145	0.99413
5	0.88560	0.97826	0.99282
6	0.90069	0.98674	0.99675
7	0.90220	0.98965	0.99803
8	0.90802	0.99175	0.99861
9	0.91230	0.99368	0.99917
10	0.91461	0.99483	0.99946
11	0.91687	0.99556	0.99960
12	0.91886	0.99626	0.99973
13	0.92022	0.99678	0.99980
14	0.92156	0.99714	0.99984
15	0.92291	0.99748	0.99988
16	0.92384	0.99770	0.99991
17	0.92447	0.99796	0.99993
18	0.92534	0.99813	0.99994
19	0.92629	0.99828	0.99995
20	0.92691	0.99841	0.99996
21	0.92720	0.99853	0.99997
22	0.92780	0.99864	0.99998
23	0.92817	0.99871	0.99998
24	0.92889	0.99879	0.99998
25	0.92916	0.99883	0.99998
26	0.92945	0.99892	0.99999
27	0.92979	0.99897	0.99999
28	0.93006	0.99902	0.99999
29	0.93029	0.99907	0.99999
30	0.93051	0.99909	0.99999

2.5 MedShift

$$\mathrm{MedShift}(\mathbf{x}, \mathbf{y}) = \underset{1 \le i \le n,\ 1 \le j \le m}{\mathrm{Median}} (x_i - y_j)$$

Practical Recommendations

MedShift provides an initial insight into the absolute difference between elements of two samples. Interpretation: half of $x_i - y_j$ is smaller than $\mathrm{MedShift}(\mathbf{x}, \mathbf{y})$, the other half is larger. For samples with small Volatility, $\mathrm{MedShift}(\mathbf{x}, \mathbf{y})$ approximates pairwise differences $x_i - y_j$.

Key Facts

- Measures the median absolute difference between elements of two samples
- Domain: any real numbers
- Equals the *Hodges-Lehmann estimator* for two samples ([HL63])

Properties

$$\operatorname{MedShift}(\mathbf{x} + k_x, \mathbf{y} + k_y) = \operatorname{MedShift}(\mathbf{x}, \mathbf{y}) + k_x - k_y$$

$$\operatorname{MedShift}(k \cdot \mathbf{x}, k \cdot \mathbf{y}) = k \cdot \operatorname{MedShift}(\mathbf{x}, \mathbf{y})$$

$$\operatorname{MedShift}(\mathbf{x}, \mathbf{y}) = -\operatorname{MedShift}(\mathbf{y}, \mathbf{x})$$

2.6 MedRatio

$$\operatorname{MedRatio}(\mathbf{x}, \mathbf{y}) = \underset{1 \le i \le n,\ 1 \le j \le m}{\operatorname{Median}} \left(\frac{x_i}{y_j} \right)$$

Practical Recommendations

MedRatio provides an initial insight into the ratio between elements of two samples expressed as a multiplicative factor. It answers “how many times larger is **x** compared to **y**?” E.g., MedRatio = 2.0 means that for 50% of pairs (x_i, y_j), x_i is at least twice as large as y_j.

MedRatio functions as a division operator: $\operatorname{MedRatio}(\mathbf{x}, \mathbf{y})$ computes the typical ratio $\mathbf{x}/\mathbf{y}$. Results convert to percentage differences as $(\operatorname{MedRatio} - 1) \times 100\%$.

Key Facts

- Measures the median ratio between elements of two samples
- Domain: $y_j > 0$
- Second sample y is always the baseline
- In general, $\operatorname{MedRatio}(\mathbf{x}, \mathbf{y}) \neq 1/\operatorname{MedRatio}(\mathbf{y}, \mathbf{x})$ (e.g., $x = (1, 100)$, $y = (1, 10)$)

Properties

$$\operatorname{MedRatio}(k_x \cdot \mathbf{x}, k_y \cdot \mathbf{y}) = \frac{k_x}{k_y} \cdot \operatorname{MedRatio}(\mathbf{x}, \mathbf{y})$$

2.7 MedSpread

$$\operatorname{MedSpread}(\mathbf{x}, \mathbf{y}) = \frac{n \operatorname{Spread}(\mathbf{x}) + m \operatorname{Spread}(\mathbf{y})}{n + m}$$

Practical Recommendations

MedSpread primarily serves as a scaling factor for MedDisparity. It represents the combined dispersion of both samples, weighted by sample size. Works best for distributions with similar dispersion values.

Key Facts

- Measures average dispersion across two samples
- Domain: any real numbers
- Provides a robust alternative to the pooled standard deviation

Properties

$$\operatorname{MedSpread}(\mathbf{x}, \mathbf{x}) = \operatorname{Spread}(\mathbf{x})$$

$$\operatorname{MedSpread}(k_1 \cdot \mathbf{x}, k_2 \cdot \mathbf{x}) = \frac{|k_1| + |k_2|}{2} \cdot \operatorname{Spread}(\mathbf{x})$$

$$\operatorname{MedSpread}(\mathbf{x}, \mathbf{y}) = \operatorname{MedSpread}(\mathbf{y}, \mathbf{x})$$

$$\operatorname{MedSpread}(k \cdot \mathbf{x}, k \cdot \mathbf{y}) = |k| \cdot \operatorname{MedSpread}(\mathbf{x}, \mathbf{y})$$

2.8 MedDisparity

$$\operatorname{MedDisparity}(\mathbf{x}, \mathbf{y}) = \frac{\operatorname{MedShift}(\mathbf{x}, \mathbf{y})}{\operatorname{MedSpread}(\mathbf{x}, \mathbf{y})}$$

Practical Recommendations

MedDisparity provides a scale-invariant insight into the absolute difference between elements of two samples, expressed in standardized spread units.

Key Facts

- Measures a normalized absolute difference between $\mathbf{x}$ and $\mathbf{y}$ expressed in standardized spread units
- Domain: $\operatorname{MedSpread}(\mathbf{x}, \mathbf{y}) > 0$ (at least 50% of $|x_i - x_j|$ and 50% of $|y_i - y_j|$ are non-zeros)
- Expresses the *effect size*, renamed to “Disparity” for clarity
- Scale-invariant, which makes an experiment design more portable

Comparison

- Compared to *Cohen’s d*: more robust while maintaining efficiency under normality

Properties

$$\operatorname{MedDisparity}(\mathbf{x} + k, \mathbf{y} + k) = \operatorname{MedDisparity}(\mathbf{x}, \mathbf{y})$$

$$\operatorname{MedDisparity}(k \cdot \mathbf{x}, k \cdot \mathbf{y}) = \operatorname{sign}(k) \cdot \operatorname{MedDisparity}(\mathbf{x}, \mathbf{y})$$

$$\operatorname{MedDisparity}(\mathbf{x}, \mathbf{y}) = -\operatorname{MedDisparity}(\mathbf{y}, \mathbf{x})$$

3 Studies

This section analyzes the estimators’ properties using mathematical proofs and Monte Carlo simulations. Most proofs are adopted from various textbooks and papers, but only the most essential references are provided.

Unlike the main part of the manual, studies require knowledge of classic statistical methods. Well-known facts and commonly accepted notation are used without special introduction. The studies provide a deep dive into properties of considered estimators for practitioners interested in rigorous proofs and results of numerical simulations.

3.1 Asymptotic Gaussian Expected Value of the Spread

This study establishes that Spread has expected value $\sqrt{2}\,\Phi^{-1}(0.75) \approx 0.954$ under standard normal data as sample size increases.

The key insight is that pairwise absolute differences $|X_i - X_j|$ from normal data converge to a known distribution. Since Spread takes the median of these differences, its asymptotic expectation equals the population median of $|X_1 - X_2|$ where $X_1, X_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

Consider $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. For any fixed $i \neq j$, the difference $X_i - X_j$ has mean 0 and variance

$$\mathbb{V}[X_i - X_j] = \mathbb{V}[X_i] + \mathbb{V}[X_j] = 1 + 1 = 2$$

because X_i and X_j are independent. Therefore

$$X_i - X_j \sim \mathcal{N}(0,2)$$

Define $W = (X_i - X_j)/\sqrt{2} \sim \mathcal{N}(0,1)$ and $D = |X_i - X_j| = \sqrt{2}\,|W|$. The cumulative distribution function of D for $t \geq 0$ is

$$\mathbb{P}[D \leq t] = \mathbb{P}\big[|W| \leq t/\sqrt{2}\big] = 2\Phi\big(t/\sqrt{2}\big) - 1$$

where Φ denotes the standard normal cumulative distribution function. The population median m of D satisfies

$$2\Phi\big(m/\sqrt{2}\big) - 1 = \tfrac{1}{2}$$

Solving for m yields

$$m = \sqrt{2}\,\Phi^{-1}(0.75)$$

The multiset of gaps $\{|X_i - X_j| : i < j\}$ forms a bounded-kernel U-statistic of degree 2. U-statistic consistency results imply that its empirical distribution converges almost surely to the law of D. The sample median of pairwise differences thus converges in probability to m. Convergence in probability plus uniform integrability yields convergence of expectations:

$$\lim_{n\to\infty} \mathbb{E}\big[\text{Spread}(X_1, \ldots, X_n)\big] = \sqrt{2}\,\Phi^{-1}(0.75)$$

or

$$\lim_{n\to\infty} \mathbb{E}\big[\text{Spread}(X_1, \ldots, X_n)\big] \approx 0.953\,873$$

3.2 Asymptotic Gaussian Efficiency of the Center

This study shows that Center achieves $3/\pi \approx 95.5\%$ efficiency relative to the sample mean under normality. This high efficiency allows Center to handle outliers while maintaining near-optimal performance on normal data.

The analysis uses U-statistic theory to derive the asymptotic distribution of Center under Gaussian data. Let

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad n \geq 2,\ \sigma > 0$$

Center($\mathbf{x}$) has translation invariance, so setting $\mu = 0$ preserves generality. The Walsh averages are

$$W_{ij} = \frac{X_i + X_j}{2}, \qquad 1 \leq i \leq j \leq n$$

Since $X_i + X_j \sim \mathcal{N}(0, 2\sigma^2)$, it follows that

$$W_{ij} \sim \mathcal{N}\big(0, \sigma^2/2\big), \qquad f_W(0) = \frac{1}{\sigma\sqrt{\pi}}$$

The estimator Center($\mathbf{x}$) equals the sample median of the $\binom{n+1}{2}$ Walsh averages. As a U-quantile of degree two, it satisfies

$$0 = \frac{2}{n(n+1)} \sum_{1 \leq i \leq j \leq n} \Big\{ \mathbf{1}\{W_{ij} \leq \text{Center}(\mathbf{x})\} - \tfrac{1}{2} \Big\}$$

U-quantile theory [Sen63] provides the linear expansion

$$\sqrt{n}\ \text{Center}(\mathbf{x}) = \frac{2}{f_W(0)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i) + o_{\mathbb{P}}(1)$$

where

$$\psi(x) = \mathbb{P}\Big\{ \tfrac{x + X_2}{2} \leq 0 \Big\} - \tfrac{1}{2} = \tfrac{1}{2} - \Phi\Big(\tfrac{x}{\sigma}\Big)$$

and Φ is the standard normal cumulative distribution function. Since $X_1/\sigma \sim \mathcal{N}(0, 1)$,

$$\mathbb{V}\big[\psi(X_1)\big] = \int_{-\infty}^{\infty} \Big(\tfrac{1}{2} - \Phi(u) \Big)^2 \varphi(u)\, du = \frac{1}{12}$$

with φ the standard normal probability density function. Substituting $\mathbb{V}[\psi(X_1)]$ and $f_W(0)$ yields

$$\mathbb{V}\big[\sqrt{n}\ \text{Center}(\mathbf{x})\big] = \frac{4 \cdot \frac{1}{12}}{\big(1/(\sigma\sqrt{\pi})\big)^2} = \frac{\pi\sigma^2}{3}$$

so

$$\sqrt{n}\ \text{Center}(\mathbf{x}) \xrightarrow{d} \mathcal{N}\Big(0, \tfrac{\pi\sigma^2}{3}\Big), \qquad n \to \infty$$

The sample mean has asymptotic variance σ^2/n, hence the *asymptotic Gaussian efficiency* of Center is

$$\text{eff}_{\mathcal{N},\infty}(\text{Center}) = \frac{\sigma^2/n}{\pi\sigma^2/(3n)} = \frac{3}{\pi} \approx 0.954\,930$$

Thus matching the mean's precision under normality requires only $1/0.955 \approx 1.05$ times as many observations, a negligible price for the Center's 29% breakdown point and much stronger resistance to outliers.

3.3 Asymptotic Gaussian Efficiency of the Median

This study shows that the sample median achieves only $2/\pi \approx 63.7\%$ efficiency relative to the sample mean under normality.

The analysis applies classical asymptotic theory for sample quantiles to derive the limiting distribution of the median under Gaussian data. Consider

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad n \geq 2,\ \sigma > 0$$

Since the sample median is translation invariant, set $\mu = 0$ without loss of generality. Write

$$M_n = \text{Median}(X_1, \ldots, X_n)$$

For any continuous distribution with density f positive at its median θ, classical theory ([SSH99], [Ser09]) gives

$$\sqrt{n}\,(M_n - \theta) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{4f(\theta)^2}\Big)$$

In the normal case $\theta = 0$ and

$$f(0) = \frac{1}{\sigma\sqrt{2\pi}}$$

so the asymptotic variance becomes

$$\frac{1}{4f(0)^2} = \frac{1}{4}\left(\sigma\sqrt{2\pi}\right)^2 = \frac{\pi\sigma^2}{2}$$

Hence

$$\sqrt{n}\,M_n \xrightarrow{d} \mathcal{N}\Big(0, \frac{\pi\sigma^2}{2}\Big), \qquad n \to \infty$$

The sample mean has asymptotic variance σ^2/n, so the *asymptotic Gaussian efficiency* of the median is

$$\text{eff}_{\mathcal{N},\infty}(\text{Median}) = \frac{\sigma^2/n}{\pi\sigma^2/(2n)} = \frac{2}{\pi} \approx 0.637$$

Thus achieving the same precision as the mean under normality requires roughly $1/0.637 \approx 1.57$ times as many observations when using the median. This large efficiency loss shows why we prefer the Hodges–Lehmann Center estimator — which attains about 95.5% efficiency — whenever data are roughly Gaussian but may include outliers.

3.4 Asymptotic Gaussian Efficiency of the Spread

This study shows that Spread achieves approximately 86% efficiency relative to the sample standard deviation under normality.

The analysis uses U-statistic theory to derive the asymptotic distribution of Spread as a scale estimator under Gaussian data. Consider

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad n \geq 2,\ \sigma > 0$$

Since $\text{Spread}(\mathbf{x}) = \text{Median}\,|X_i - X_j|$ is translation invariant, set $\mu = 0$ without loss of generality and write

$$R_n = \text{Spread}(\mathbf{x})$$

$$m_0 = \sqrt{2}\,\Phi^{-1}(0.75)\,\sigma \approx 0.954\,\sigma$$

Here m_0 denotes the population median of $|X_i - X_j|$ when $X_i \sim \mathcal{N}(0, \sigma^2)$. Letting $D = |X_1 - X_2|$, its density is

$$f_D(t) = \frac{1}{\sigma\sqrt{\pi}} \exp\Big(-\frac{t^2}{4\sigma^2}\Big), \qquad t \geq 0$$

with $f_D(m_0) = \sigma^{-1}\pi^{-1/2}\exp(-\Phi^{-1}(0.75)^2/2)$. Treating R_n as a degree-two U-quantile and applying asymptotic theory [Sen63] yields

$$\sqrt{n}\,\big(R_n - m_0\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{4\,\mathbb{V}[\psi(X_1)]}{f_D(m_0)^2}\Big)$$

where

$$\begin{aligned}\psi(x) &= \mathbb{P}\big\{|x - X_2| \leq m_0\big\} - \tfrac{1}{2} \\ &= \Phi\Big(\tfrac{x+m_0}{\sigma}\Big) - \Phi\Big(\tfrac{x-m_0}{\sigma}\Big) - \tfrac{1}{2}\end{aligned}$$

Numerical evaluation of the integral

$$\mathbb{V}[\psi(X_1)] = \int_{-\infty}^{\infty} \psi(x)^2\, \frac{e^{-x^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}\, dx$$

yields $\mathbb{V}[\psi(X_1)] \approx 0.0266$. Substituting this value and $f_D(m_0)$ into the variance formula gives

$$\mathbb{V}\big[\sqrt{n}\,R_n\big] \approx 0.527\,\sigma^2$$

Because R_n is *not* a consistent estimator of σ, comparisons with the sample standard deviation $\text{StdDev}(\mathbf{x})$ use the rescaled statistic

$$\widehat{\sigma}_{\text{Spread}} = \frac{R_n}{m_0}$$

which *is* consistent. Dividing the variance above by the constant m_0^2 gives

$$\mathbb{V}\big[\sqrt{n}\,\widehat{\sigma}_{\text{Spread}}\big] \approx 0.579\,\sigma^2$$

The optimal Gaussian scale estimator $\mathrm{StdDev}(\mathbf{x})$ has asymptotic variance $\sigma^2/(2n)$, so the *asymptotic Gaussian efficiency* of the (scaled) Spread is

$$\mathrm{eff}_{\mathcal{N},\infty}(\mathrm{Spread}) = \frac{\sigma^2/(2n)}{0.579\,\sigma^2/n} \approx 0.864$$

One needs roughly $1/0.864 \approx 1.16$ times as many observations to match the precision of the sample standard deviation when the data are exactly normal. In exchange, Spread inherits a 29% breakdown point from its U-quantile construction, so moderate extra data provide a substantial increase in robustness.

3.5 Asymptotic Gaussian Efficiency of the Median Absolute Deviation

This study shows that the median absolute deviation (MAD) achieves only about 37% efficiency relative to the sample standard deviation under normality.

The analysis derives the asymptotic distribution of the MAD under Gaussian data using classical theory for sample medians. Consider

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad n \geq 2,\ \sigma > 0$$

Since both the sample median and the MAD are translation invariant, set $\mu = 0$ without loss of generality and write

$$\mathrm{MAD}_n = \mathrm{Median}\Big(|X_i - \mathrm{Median}(\mathbf{x})| : i = 1, \ldots, n\Big)$$

For a standard normal distribution, the population MAD equals

$$m_0 = \Phi^{-1}(0.75)\,\sigma = c_0\,\sigma$$

$$c_0 = \Phi^{-1}(0.75) \approx 0.674$$

Dividing the empirical MAD by this constant gives a consistent scale estimator:

$$\widehat{\sigma}_{\text{MAD}} = \frac{\mathrm{MAD}_n}{c_0}$$

To find its large-sample variance, observe that $Y_i = |X_i|$ has density

$$g(y) = \frac{2}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{y^2}{2\sigma^2}\Big), \qquad y \geq 0$$

whose median is m_0. A classical result for sample medians of independent draws with continuous positive density at the median ([SSH99], [Ser09]) states

$$\sqrt{n}\big(\mathrm{Median}(Y_1, \ldots, Y_n) - m_0\big) \xrightarrow{d} \mathcal{N}\Big(0, \tfrac{1}{4g(m_0)^2}\Big)$$

Since

$$g(m_0) = \frac{2}{\sigma\sqrt{2\pi}} \exp\Bigl(-\frac{c_0^2}{2}\Bigr)$$

the asymptotic variance of $\sqrt{n}\ \mathrm{MAD}_n$ is

$$\mathbb{V}\bigl[\sqrt{n}\ \mathrm{MAD}_n\bigr] = \frac{1}{4g(m_0)^2} = \frac{\pi\sigma^2 \exp(c_0^2)}{8}$$

Scaling by $1/c_0$ gives the variance of the consistent estimator:

$$\mathbb{V}\bigl[\sqrt{n}\,\widehat{\sigma}_{\mathrm{MAD}}\bigr] = \frac{\pi\sigma^2 \exp(c_0^2)}{8\,c_0^2}$$

The optimal Gaussian scale estimator is the sample standard deviation. It has asymptotic variance of $\sigma^2/(2n)$. The *asymptotic Gaussian efficiency* of the (scaled) MAD is therefore

$$\mathrm{eff}_{\mathcal{N},\infty}(\mathrm{MAD}) = \frac{\sigma^2/(2n)}{\pi\sigma^2 \exp(c_0^2)/(8c_0^2 n)} = \frac{4c_0^2}{\pi\ \exp(c_0^2)}$$

This gives the result:

$$\mathrm{eff}_{\mathcal{N},\infty}(\mathrm{MAD}) \approx 0.367\,523$$

Achieving the same precision as the sample standard deviation under normality requires roughly $1/0.368 \approx 2.7$ times as many observations when using the MAD.

3.6 Finite-Sample Efficiency of Central Tendency Estimators

This study presents finite-sample efficiency values for Center and shows how it performs better than Median across small and medium sample sizes.

The previous studies established asymptotic efficiency values — the limiting behavior as sample size approaches infinity. For the Gaussian distribution, these asymptotic values are:

- Mean: 100% (the most efficient estimator under normality)
- Center: $3/\pi \approx 95.5\%$
- Median: $2/\pi \approx 63.7\%$

Asymptotic theory provides excellent approximations for large samples but may be inaccurate for small n. Finite-sample efficiency captures the actual precision when working with limited data.

Efficiency and Sample Size Requirements

Efficiency quantifies the variance ratio between estimators. For two estimators T_1 and T_2 applied to the same distribution:

$$\mathrm{eff}(T_1 \text{ relative to } T_2) = \frac{\mathbb{V}[T_2]}{\mathbb{V}[T_1]}$$

This ratio directly translates to sample size requirements. An estimator with 80% efficiency needs $100/80 = 1.25$ times as many observations to achieve the same precision as the reference estimator. The Median with its 63.7% asymptotic efficiency requires roughly 1.57 times more data than the Mean under normality.

Monte Carlo Estimation of Finite-Sample Efficiency

Numerical simulation provides exact efficiency values for any sample size. The procedure follows these steps:

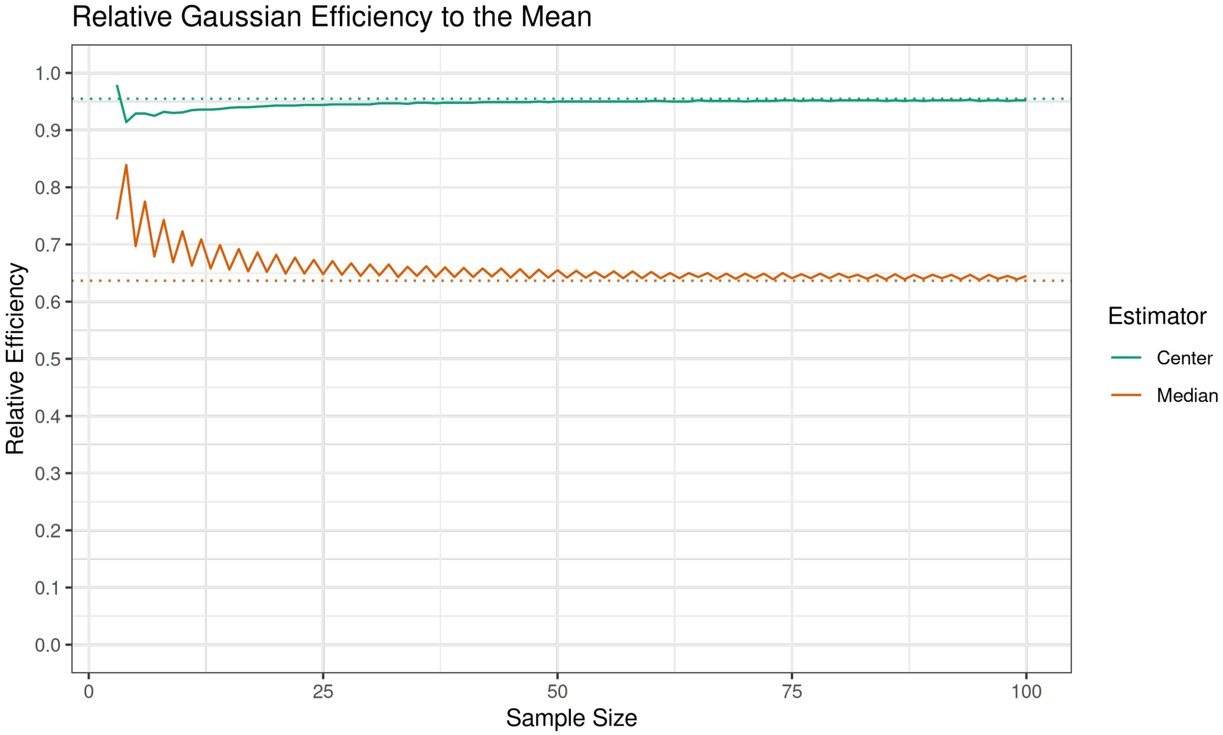
1. **Generate samples**: Draw m independent samples of size n from the standard normal distribution
2. **Calculate estimators**: Compute both estimators for each sample
3. **Measure dispersion**: Calculate the sample variance of the m estimator values
4. **Compute efficiency**: Take the variance ratio

The simulation must balance computational cost against precision. Larger m reduces Monte Carlo error but increases computation time. For efficiency estimation, $m = 10^6$ iterations achieve enough precision.

Finite-Sample Results

The simulation reveals how efficiency evolves from small to moderate sample sizes.

The figure below shows the Gaussian efficiency curves for $n \in \{3, \dots, 100\}$ based on $m = 10^6$ Monte Carlo iterations (dotted lines show asymptotic values):



Key Observations

1. **Small sample behavior**: For $n = 3$, Center achieves 97.9% efficiency while Median drops to 74.3%. Even at extreme small samples, Center maintains above 91% efficiency throughout.

2. **Convergence patterns**: The Center estimator reaches 94% efficiency by $n = 15$ and stabilizes above 95% for $n \geq 50$. The Median oscillates between 63.5% and 68% across all sample sizes, never approaching the efficiency of Center.

3. **Practical advantage**: For typical applications ($n = 20$ to 50), Center maintains $94 - 95\%$ efficiency. This represents only a $5 - 6\%$ penalty compared to Mean, while Median suffers a $35 - 37\%$ penalty.

The toolkit's Center estimator performs much better than Median while sacrificing minimal efficiency compared to the non-robust Mean.

3.7 Finite-Sample Efficiency of Dispersion Estimators

This study presents finite-sample efficiency values for Spread and demonstrates its consistent superiority over the *standard deviation* StdDev and the *median absolute deviation* MAD across small and medium sample sizes.

The previous studies established asymptotic efficiency values for dispersion estimators under the Gaussian distribution. These asymptotic values are:

- StdDev: 100% (the most efficient estimator under normality)
- Spread: $\approx 86.4\%$
- MAD: $\approx 36.8\%$

The asymptotic values provide excellent approximations for large samples but may not accurately reflect performance with limited data. Understanding actual efficiency at practical sample sizes guides estimator selection from the toolkit.

Efficiency Measurement for Dispersion Estimators

Dispersion estimators face a variance-bias trade-off: different estimators have different expected values under the same distribution. To compare their variability fairly, the simulation normalizes each estimator by its expected value. This normalization eliminates bias differences and focuses purely on variance comparison.

The procedure follows these steps:

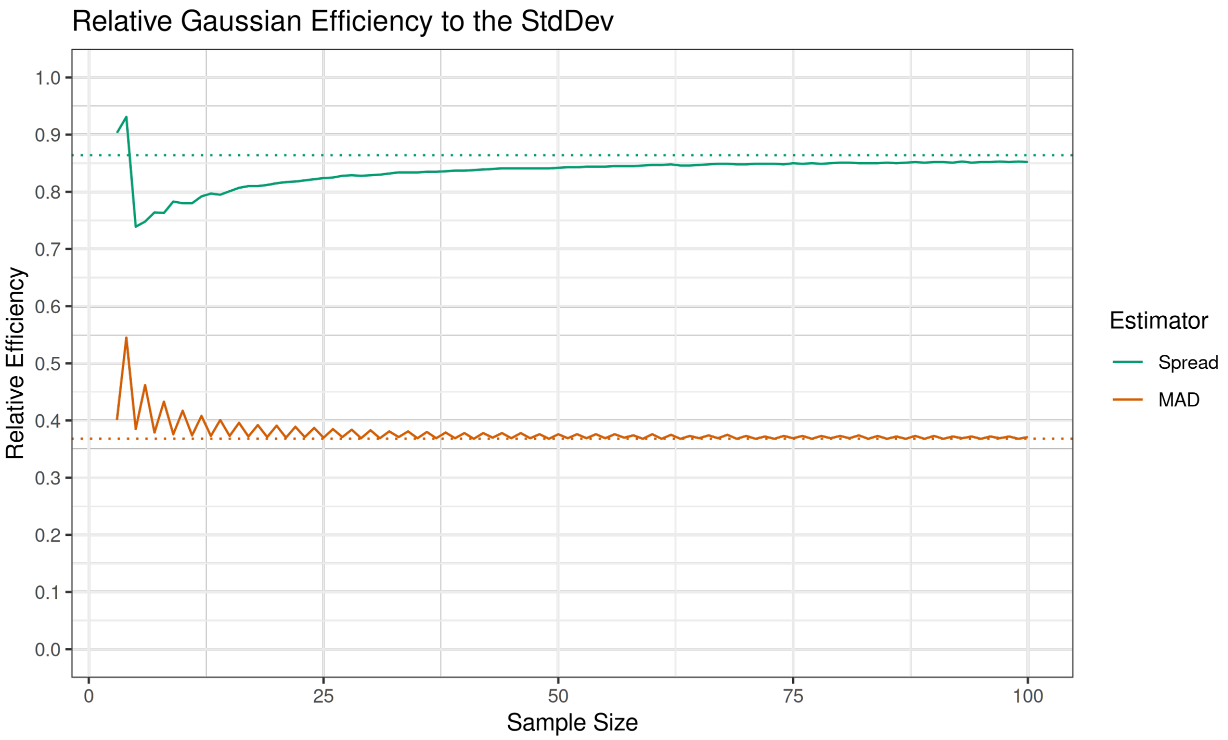
1. **Generate samples**: Draw m independent samples of size n from the standard normal distribution
2. **Calculate estimators**: Compute StdDev, Spread, and MAD for each sample
3. **Calculate normalized variance**: for each set of estimations, calculate the variance and divide by the mean value to align estimator biases
4. **Compute efficiency**: Take the variance ratio relative to StdDev

This approach ensures fair comparison by measuring how much each estimator varies around its own expected value. The simulation uses $m = 10^6$ iterations to achieve sufficient precision.

Results and Analysis

The simulation covers sample sizes $n \in \{3, \ldots, 100\}$ with efficiency computed relative to standard deviation.

The below figure shows the Gaussian efficiency curves based on 10^6 Monte Carlo iterations (dotted lines show asymptotic values):



Key Observations

1. **Superior efficiency**: Spread consistently outperforms MAD by a factor of two across all sample sizes. At $n = 20$, Spread achieves 81.4% efficiency while MAD reaches only 39.0%.
2. **Practical implications**: For typical sample sizes ($n = 20$ to 50), Spread maintains $81 - 84\%$ efficiency. This $16 - 19\%$ penalty compared to StdDev represents a reasonable trade-off for robustness.
3. **Small sample behavior**: Spread shows a characteristic dip for $n = 5 - 8$ (dropping to 74%) before recovering. Even at its worst, Spread remains twice as efficient as MAD.

Spread from the toolkit provides a substantial efficiency advantage over MAD while maintaining reasonable performance compared to the non-robust standard deviation. This efficiency advantage, combined with its robustness properties, makes Spread the preferred choice for dispersion estimation in practical applications.

References

[HL63] J. L. Hodges and E. L. Lehmann. “Estimates of Location Based on Rank Tests.” en. In: *The Annals of Mathematical Statistics* 34.2 (June 1963), pp. 598–611. ISSN: 0003-4851. DOI: 10.1214/aoms/1177704172. URL: http://projecteuclid.org/euclid.aoms/1177704172.

[Sen63] Pranab Kumar Sen. “On the Estimation of Relative Potency in Dilution (-Direct) Assays by Distribution-Free Methods.” In: *Biometrics* 19.4 (Dec. 1963), p. 532. ISSN: 0006341X. DOI: 10.2307/2527532. URL: https://www.jstor.org/stable/2527532.

[Ser09] Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

[Sha76] Michael Ian Shamos. “Geometry and Statistics: Problems at the Interface.” In: (1976).

[SSH99] Zbynek Sidak, Pranab Kumar Sen, and Jaroslav Hajek. *Theory of rank tests*. Elsevier, 1999.