# A Simulation Study of Robust Statistics

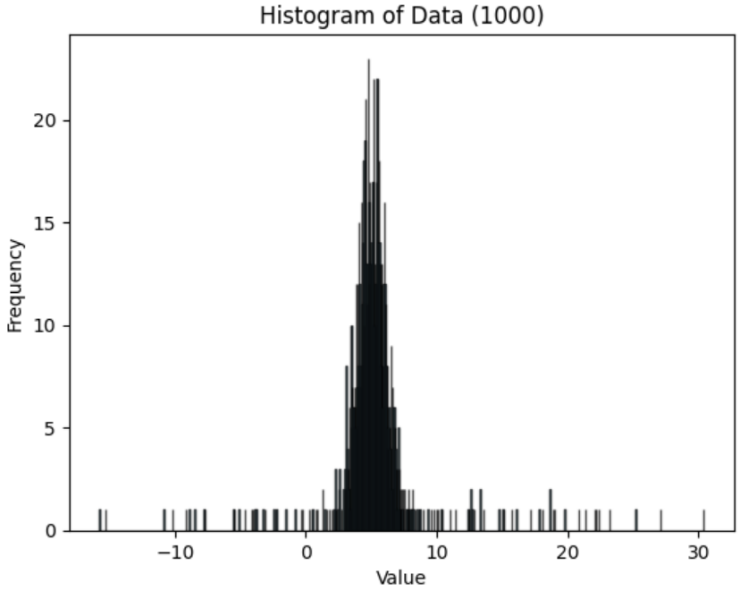
by Anthony Xing
Mentor: Ethan Ancell

# What I Did This Quarter

Throughout this quarter, we learned many mathematical and theoretical ideas from robust statistics, but at the end we also worked on a simulation study. The other mentee Yu He Zhang is presenting on the theoretical parts of this project, and I will be presenting on the simulation study here.

# What should Robust Statistics Achieve?

Robust statistics refer to statistical methods and measures that remain effective and reliable even when the underlying assumptions about the data are violated, examples being the presence of outliers or deviations from a specified distributional model.



# Simulation Study: Estimates of Mu

Let's consider the following estimates of the parameter $\mu$:

- **Mean:** $\bar{X}$
- **Median:** $X^{(\text{med})}$
- **$\alpha$-trimmed mean:** $\bar{X}_\alpha^{(\text{trim})}$ *(Only consider the data points from the $\alpha$ quantile to the $1-\alpha$ quantile, and then take the mean of those.)*
- **Midrange:** $X^{(\text{midrange})} = \frac{\max_i X_i - \min_i X_i}{2}$

alpha = 0.10 for alpha-trimmed mean.

Which unbiased estimator is best (most robust) for estimating the true Mu?

# Simulation Study: Sampling Data

Consider the following distribution:

$$F_\epsilon := (1-\epsilon)F_0 + \epsilon G_\rho, \tag{1}$$

where $F_0$ is the CDF of a $N(\mu, \sigma^2)$ distribution, and $G_\rho$ is the CDF of a $N(\mu, \rho^2\sigma^2)$ distribution. Then, we will be simulating data

$$X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_\epsilon. \tag{2}$$

(true) mu = 5, sigma = 1.

We'll simulate from many different values for epsilon {0.000, 0.005, 0.020, 0.100} and rho {2, 10}.

(higher epsilon → more outliers, higher rho → outliers more extreme)

F_e → has probability 1-e to sample from F_0, probability e to sample from G_p

# Simulation Study: The Process

- For 100 replications, we sample 1000 data points from F_e.

- For each replication, use the 1000 data points to obtain a value for each estimate of Mu (mean, median, a-trimmed mean, midrange)

- With 100 replications, we then have 100 values for each estimate of Mu.
- Using these values, calculate the expectation and variance for each estimate of Mu (mean, median, a-trimmed mean, midrange)

# Simulation Study: The Process

With the expectation and variance for each estimate of Mu (mean, median, a-trimmed mean, midrange), we can see how close the estimate of Mu is to the true Mu (5) and how much the estimate of Mu can vary.

Expectation of estimate of Mu being close to the true Mu is good! (accurate!)

Lower variance for an estimate of Mu is good!

Lower Variance → Efficient

Efficient even in contaminated distributions → Robust

# Simulation Study: Results for rho=2

Results for rho=2

| | Epsilon | Expectation Mean | Expectation Median | Expectation Trimmed Mean | Expectation Midrange |
|---|---|---|---|---|---|
| **0** | 0.000 | 5.000479 | 4.999382 | 5.000999 | 3.272719 |
| **1** | 0.005 | 4.998499 | 4.999548 | 4.999192 | 5.235882 |
| **2** | 0.020 | 5.003452 | 5.001453 | 5.004373 | 7.554040 |
| **3** | 0.100 | 4.998449 | 4.999307 | 4.997424 | 10.053094 |

Results for rho=2

| | Epsilon | Variance Mean | Variance Median | Variance Trimmed Mean | Variance Midrange |
|---|---|---|---|---|---|
| **0** | 0.000 | 0.001152 | 0.001692 | 0.001175 | 0.076226 |
| **1** | 0.005 | 0.001072 | 0.001668 | 0.001044 | 2.260723 |
| **2** | 0.020 | 0.001346 | 0.001759 | 0.001100 | 2.179307 |
| **3** | 0.100 | 0.002516 | 0.002069 | 0.001348 | 1.279757 |

# Simulation Study: Results for rho=10

Results for rho=10

| | Epsilon | Expectation Mean | Expectation Median | Expectation Trimmed Mean | Expectation Midrange |
|---|---|---|---|---|---|
| **0** | 0.000 | 4.997204 | 4.997938 | 4.998796 | 3.293120 |
| **1** | 0.005 | 4.968442 | 5.000440 | 4.997438 | 113.072172 |
| **2** | 0.020 | 5.072539 | 4.999891 | 5.000764 | 184.811161 |
| **3** | 0.100 | 4.996195 | 5.003393 | 5.002924 | 251.375988 |

Results for rho=10

| | Epsilon | Variance Mean | Variance Median | Variance Trimmed Mean | Variance Midrange |
|---|---|---|---|---|---|
| **0** | 0.000 | 0.000841 | 0.001603 | 0.000907 | 0.070353 |
| **1** | 0.005 | 0.045433 | 0.001208 | 0.000932 | 2546.902951 |
| **2** | 0.020 | 0.253419 | 0.001261 | 0.000848 | 1182.675482 |
| **3** | 0.100 | 0.979059 | 0.002062 | 0.001412 | 915.008592 |

# The End!

Any Questions?