Introduction to Robust Statistics

Name: Yu He Zhang

This project had two manatees, me and Anthony Xing. In this project, we did theory and simulation. Anthony’s presentation was on the simulation part of project, and my presentation was on the theory..

What is robust statistic?

Robust statistics are methods designed to produce reliable estimates even when data contains outliers or deviates from model assumptions.

Why it is important?

Unlike well-known estimators like the mean, which are highly sensitive to extreme values, robust methods minimize the impact of unusual data points.

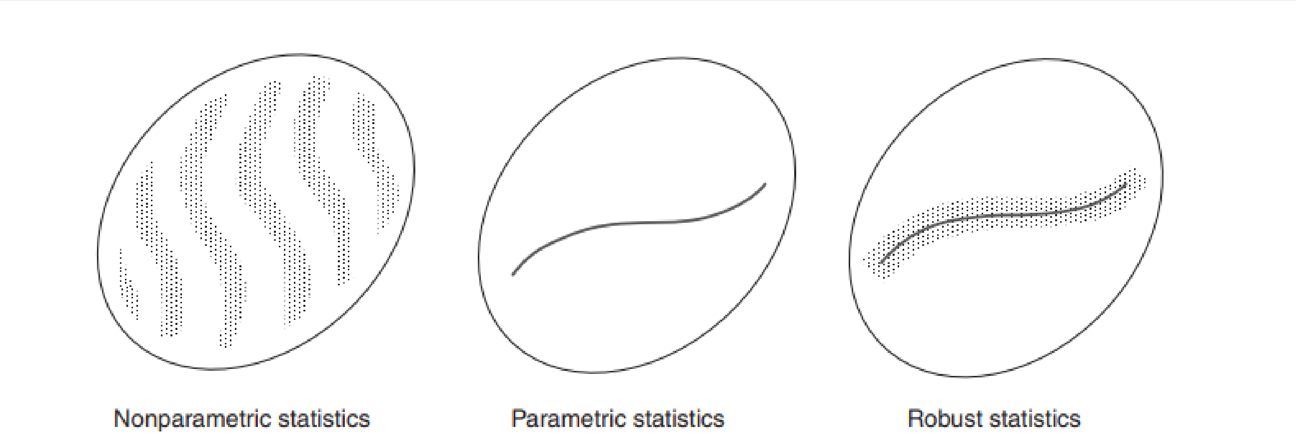
Allow estimators to not be too sensitive to outliers or misspecification.

Parametric vs. Robust vs. Nonparametric

Parametric Statistics: Assumes a specific model or distribution

Robust Statistics: Works with parametric models, but allowing for deviations and handling outliers.

Nonparametric Statistics: Assume little to nothing about the model, providing flexibility.



Key Concepts: Efficiency and Stability

Efficiency:

Efficiency measures how well an estimator uses the data. Roughly speaking, an estimator is efficient when it has low variance.

Stability:

Stability refers to how well an estimator performs when model assumptions are violated or when outliers are present.

In robust statistics, we sacrifice a small amount of efficiency to gain stability..

Test Scores Example:

- Data: 80, 85, 90, 95, and 300.
- Mean: Affected by the outlier (130).
- Median: Stable (90).
- Insight: The mean is more efficient but less stable; the median sacrifices some efficiency for greater robustness.

An Example: Mean vs. Median

Suppose that we are interested in estimating the mean/median of a symmetric distribution with the sample mean and sample median.

Sample Mean:

- In many parametric model, this is the most efficient estimator. .
- Sensitive to outliers: A single extreme value can significantly distort the result.

Sample Median:

- Less efficient: Ignores rank-order information beyond the middle point.
- Stable: Resists distortion from outliers and extreme values.

Key Insight:

- The mean is **highly efficient** but **unstable** with outliers.
- The median is **less efficient** but **more stable**, making it a robust alternative.

A Formal Definition of Robustness

Key Terms:

$T_n = T_n(X_1, X_2, \ldots, X_n)$: A statistic calculated from data

F_0 = The assumed (specified) model or distribution.

F: The actual (true) model, which may deviate from F_0

$L_F(T_n)$: The distribution of the statistic T_n under the model F.

L_{F0} : The distribution of the statistic T_n under the model F_0

d* (G_1, G_2) : A function measuring the "distance" between distributions F_1 and F_2.

A Formal Definition of Robustness

A Formal Definition of Robustness

A statistic T_n is called "**robust**" if for all $\varepsilon > 0$, there exists a $\delta > 0$ and large N such that for all $n \geq N$:

$$d^*(F, F_0) \leq \delta \implies d^*(\mathcal{L}_F(T_n), \mathcal{L}_{F_0}(T_n)) \leq \varepsilon$$

(As a reminder, $d^*(G_1, G_2)$ is how far apart distributions G_1 and G_2 are.)

Simplifying Robustness

Robust statistics ensure that **small deviations in the model lead to small deviations in the statistic**.

In simpler terms, if the true model F slightly deviates from the assumed model F_0, the statistic T_n will not deviate too much..

Robust methods prioritize stability.

This property makes robust methods valuable for messy or imperfect data, ensuring that results remain reliable even when assumptions aren't perfectly met.