Introduction to Robust Statistics

The **classical statistical** techniques are designed to be the best possible when assumptions are strictly adhered. E.g. normal errors, no outliers, homoscedasticity (var remains constant

- highly sensitive to assumption violations

Robust and resistant methods, : ‘best’ compromises for a broad range of situations and surprisingly often, are close to ‘best’ for each situation alone.

Robust statistics: focuses on developing methods that remain reliable and informative even when data deviate from idealized assumptions.

Key robust estimators

- Robust location estimates: measures of central tendency that are less sensitive to extreme values.
- Robust scales estimates: measures of dispersion that resist the influence of outliers.

What drives the need for robust statistics?

1. **Sensitivity** of statistical methods to deviations from assumptions
2. **Designing and developing statistical methods** that provide accurate and meaningful results even in the presence of outliers, heavy-tailed distributions, or model misspecifications (e.g., heavy-tailed distributions like t-distribution, presence of outliers)
3. **Quantifying** the robustness of a statistical method
4. Managing the Trade-off Between Robustness and Efficiency

For these, we introduce two key robustness measures:

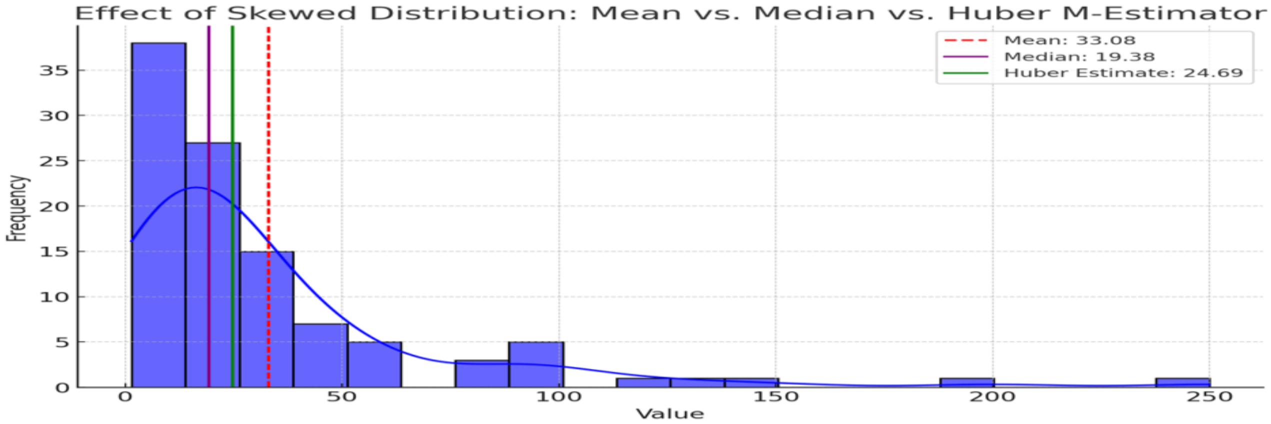
- Breakdown point
- Influence function

Impact of Robust Statistics on Statistical Analysis

Improved Reliability in Presence of Outliers

Traditional statistical estimates:

$$\left.\begin{array}{ll} \text{Mean,} & \hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \\ \text{Variance,} & \hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{array}\right\} \quad \text{sensitive to outlying observation(s)}$$

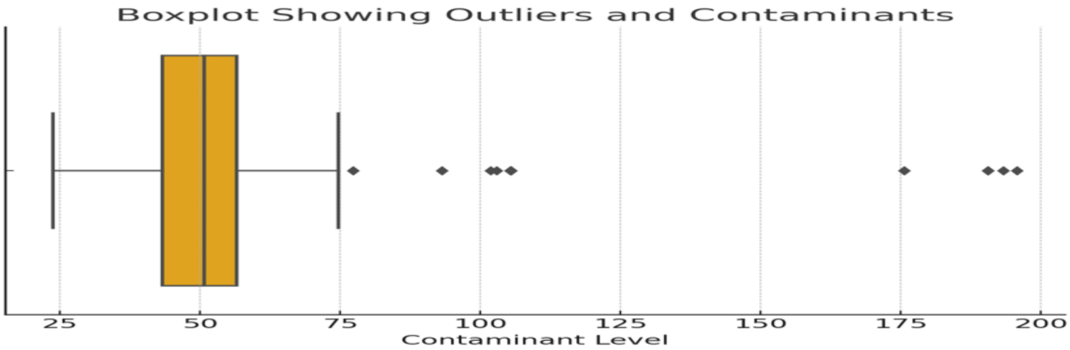


Traditional methods (like the mean and standard deviation) - severely affected by extreme values.

Robust methods, such as the median, trimmed mean, and M-estimators, provide more stable summaries of data.

Broader Applicability of Statistical Methods

Robust techniques allow statisticians to analyze such data without excessive concern about **assumption violations.**



Development of New Estimators and Algorithms

Robust statistics has led to the creation of powerful estimators like Huber's M-estimator, Tukey's biweight function, which improve estimation accuracy in the presence of non-Gaussian noise.

Enhanced Decision-Making in Uncertain Environments

Since robust statistical methods **reduce the influence of anomalies**, they enable **better decision-making**, for example:

- In finance, use robust estimators to improve risk modeling, where extreme market events can distort traditional methods.
- In medical research, apply robust statistical techniques for patient data analysis to prevent extreme values from skewing clinical findings.
- In machine learning, implement robust loss functions to **prevent overfitting** due to outliers in training data.

Course Expectations: What Students Will Gain from This Robust Statistics Course

A. **Theoretical Understanding of Robust Statistics**

- **Core Concepts & Fundamental Questions**
- **Robust Estimators and Metrics**
- **Trade-offs Between Robustness and Efficiency**

B. **Practical Skills and Applications**

- Robust Data Analysis Techniques
- Diagnostic and Model Evaluation Tools
- Robust Statistics in Real-World Applications: e.g. in finance, medicine, engineering, and machine learning.

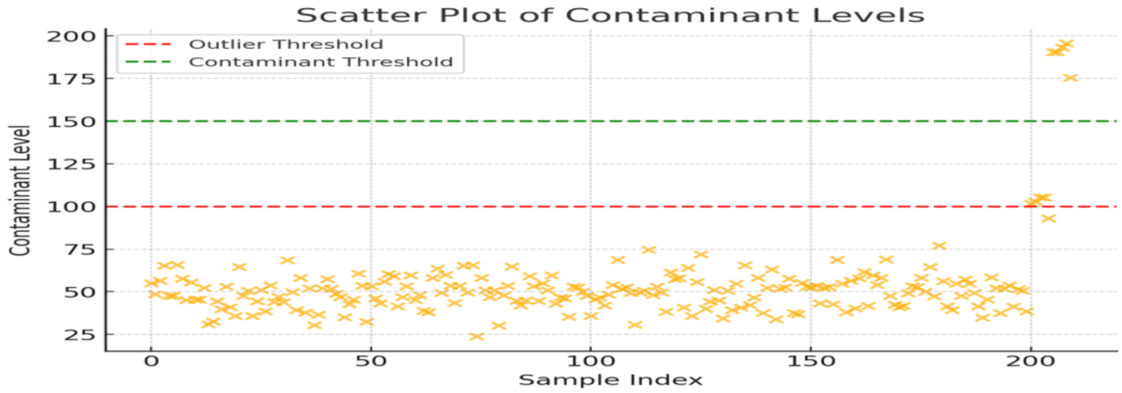
By the end of the course, students will:

✓ Be able to assess and choose appropriate statistical methods based on data conditions.
✓ Develop a strong problem-solving mindset when dealing with imperfect datasets.
✓ Gain practical experience with robust statistical software tools.
✓ Enhance research credibility by ensuring their statistical findings are not unduly influenced by outliers or model misspecifications.

- A must have for modern data analysis!

Outliers

Outliers may have arisen from purely deterministic reasons: reading, recording, or calculation error in the data. In less clear-cut circumstances, we regard this as being a random nature. An outlier in a set of data is an observation which appears to be **inconsistent** with the remainder of that set of data.



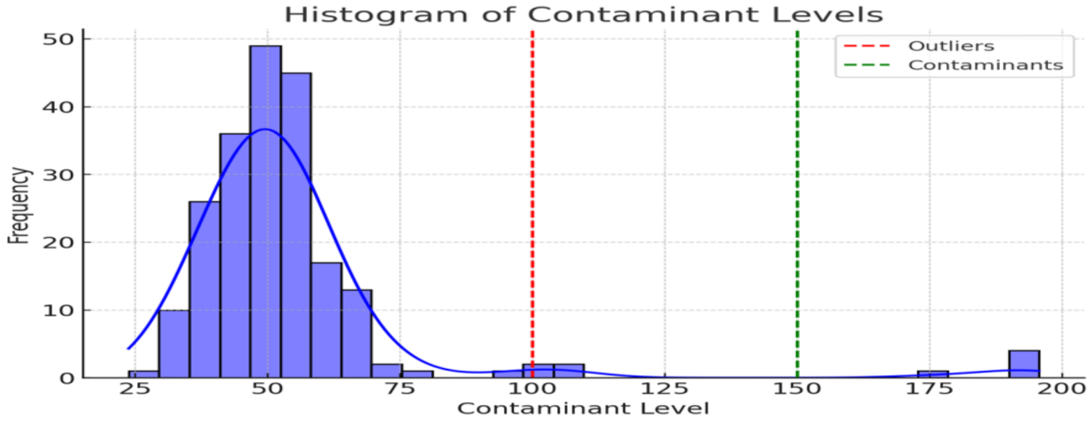
Extreme observations, outliers and contaminants

Suppose we have a random univariate sample of size n from a distribution, F: $x_1, x_2, \cdots x_n$

Then the ordered sample statistics can be written as: $x_{(1)}, x_{(2)}, \cdots x_{(n)}$

Observations $x_{(1)}$ and $x_{(n)}$ are the sample extremes.

Whether we can declare extremes as outlier depends on how they appear in relation to the postulated model, F.



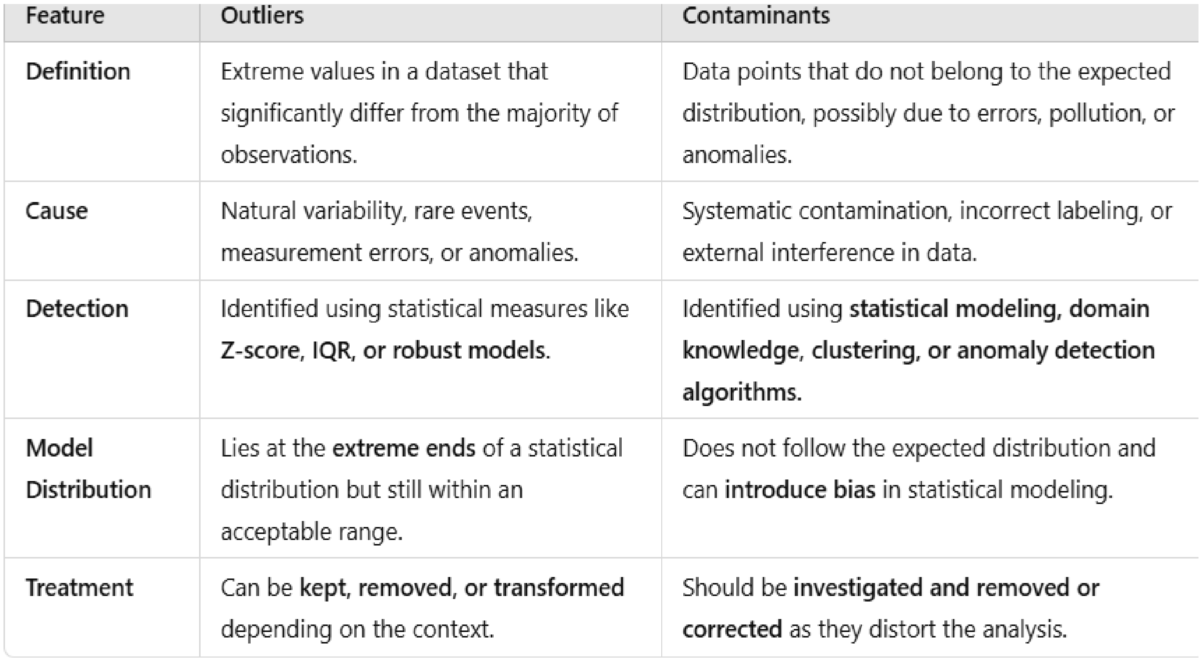
Suppose now that not all observations come from the distribution F, but one or two come from distribution G. The observations from G are termed contaminants.

Such **contaminant may appear as extremes** but need not be so.

How do we identify outliers

Z-score Method: Identifies outliers based on a threshold (|Z-score| > 3).

Clustering: Groups data based on density and labels outliers as **-1**.



Feature	Outliers	Contaminants
Definition	Extreme values in a dataset that significantly differ from the majority of observations.	Data points that do not belong to the expected distribution, possibly due to errors, pollution, or anomalies.
Cause	Natural variability, rare events, measurement errors, or anomalies.	Systematic contamination, incorrect labeling, or external interference in data.
Detection	Identified using statistical measures like **Z-score, IQR, or robust models.**	Identified using **statistical modeling, domain knowledge, clustering, or anomaly detection algorithms.**
Model Distribution	Lies at the **extreme ends** of a statistical distribution but still within an acceptable range.	Does not follow the expected distribution and can **introduce bias** in statistical modeling.
Treatment	Can be **kept, removed, or transformed** depending on the context.	Should be **investigated and removed or corrected** as they distort the analysis.

Note:

- Extreme values may or may not be outliers. Any outliers, however, are always extreme.
- Outliers may or may not be contaminants; contaminants may or may not be outliers.

Identification problem: Outliers in univariate data.

The classical summary statistics for a batch of data consisting of n observations $x_1,\cdots,x_n$ involve simple arithmetic operation on the data. Examples are

(i) sample mean: $\bar{x} = \dfrac{(x_1 + \cdots + x_n)}{n}$

(ii) sample variance: $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

(iii) residual, $r_i^2 = (X_i - \bar{X})^2,\ i = 1,\cdots,n$

Question: What happen when $x_k,\ 1 < k < n$, is large?

Identifies outliers based on a threshold (|Z-score| > 3).

Alternative measures

Let the order statistics of the sample $x_1, \cdots, x_n$ be denoted by $x_{(1)}, \cdots, x_{(n)}$, where $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$.

Define the rank of the data in either of two ways, i.e. ascending (rank upward) or descending (rank downward) order.

Definition: The **depth of a data value** generally refers to how deeply it is embedded in the data structure or how it relates to **statistical depth functions** in data analysis.

Here are some interpretations:

Depth in Statistical Analysis

- **Data depth** is a measure of how central or extreme a data point is within a dataset.
- In **robust statistics**, depth functions rank observations based on their "centrality" within the data distribution.
- Example**:** A median has the highest depth: $(n+1)/2$; while extreme values (outliers) have the lowest depth.

Depth in Outlier Detection

- This represents how "inside" or "outside" a point is relative to the distribution.
- Outliers typically have **low depth**, meaning they are distant from the majority of the data.
- Techniques like Tukey's Depth, Mahalanobis Depth, or Half-Space Depth can be used to quantify depth.

To the median and the extremes, we add another pair of summary values, the hinges, which we called fourths and is defined by

$$\text{depth of fourth} = \frac{[\text{depth of median}] + 1}{2}$$

Thus, each fourth comes halfway between the median and the corresponding extremes.

Spreads

An alternative to standard deviation is the fourth-spread, d_F, which is defined as

Fourth-spread = (upper fourth) – (lower fourth)

i.e.

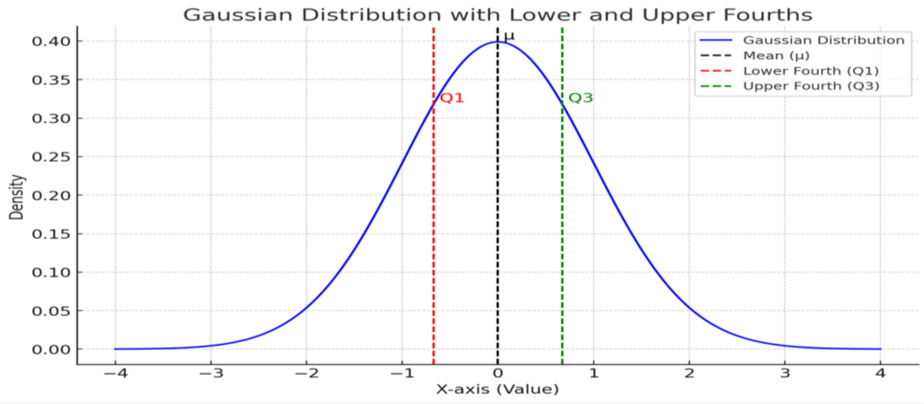
$$d_F = F_U - F_L$$

Identifying outlying values:

Any values lying outside the interval $(F_L - 1.5d_F, F_U + 1.5d_F)$, i.e. outside cutoffs, will be flagged as outlying observation.

Illustration:

Consider the case of observations selected from a Gaussian distribution.



Note:

1. The fourth corresponds to a tail of 0.25.
   Under the Gaussian distribution, the lower fourth is $F_L = \mu - 0.6745\sigma$
2. The F-spread corresponds to 1.349σ
3. The cutoff corresponds to $(\mu - 2.689\sigma, \mu + 2.689\sigma)$, thus corresponds to a tail area of 0.00347 in each tail.

Recap:

Deriving Mean, Variance, and Fourths from a Probability Density Function (PDF)

When a probability density function (PDF), $f(x)$, is given, we can derive important statistical measures such as **mean**, **variance**, and **quartiles (upper and lower fourths)** using integral calculus. These measures help in summarizing the central tendency and dispersion of continuous probability distribution.

The **mean** or **expected** $\boldsymbol{E(X)}$ of a continuous random variable X with pdf $f(x)$ is defined as:

$$\mu = (X) = \int_{-\infty}^{\infty} xf(x)dx$$

Interpretation!

Variance measures how much the values of X deviate from the mean. It is given by:

$$\sigma^2 = E[X^2] - (E[X])^2$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx$$

Thus, the variance formula becomes:

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \left(\int_{-\infty}^{\infty} xf(x)dx\right)^2$$

Interpretation:

The **standard deviation** σ is simply: $\sigma = \sqrt{\sigma^2}$

Quartiles are derived using the **cumulative distribution function (CDF)**:

$$F(X) = \int_{-\infty}^{x} f(t)dt$$

The **Lower Fourth $\boldsymbol{F_L}$** is the value of x at which **25%** of the probability mass lies to the left:

$$F(Q_1) = 0.25$$

which means solving for Q_1 $\int_{-\infty}^{Q_1} f(x)dx = 0.25$

The upper fourth???

Interpretation:

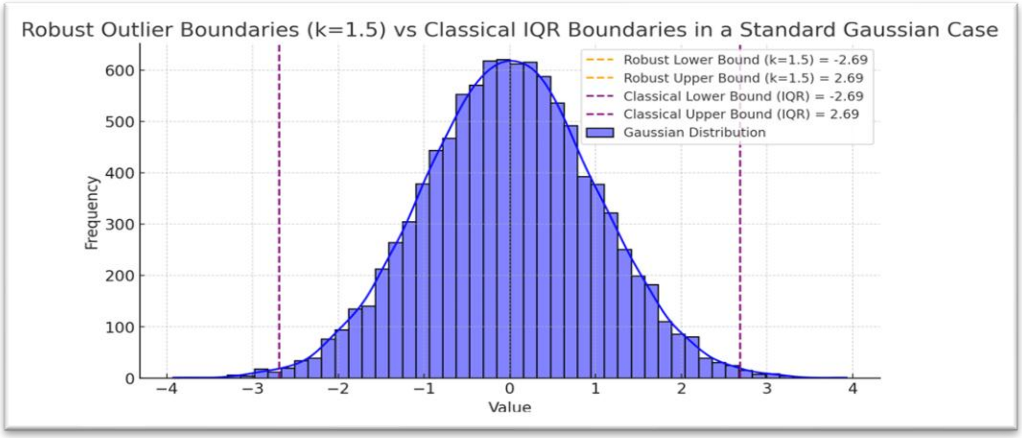
- The interquartile range (**IQR**)
- It represents the **middle 50%** of the data and is useful for detecting **outliers**.

Replacement for the standard deviation or variance.

The corresponding standard deviation in resistant measure is: $\sigma_F = \frac{d_f}{1.349}$ (F-pseudosigma)

Note: When the data are Gaussian, the F-pseudosigma yields an estimate of σ, and its value is usually close to s, the sample standard deviation.

Consider the standard Gaussian distribution, with mean 0 and variance 1.



For this symmetric distribution

- median equals the mean:
- population fourths:
- population fourth-spread: 1.349.
- population cutoffs are $\pm 2\sigma_F$

Note: Compare this with the classical method of detection. Comment on your findings.

Try some exercises!

Introduction to Boxplots and Batch Comparisons: A Practical Guide for Data Analysis & Visualization

Boxplots, also known as box-and-whisker plots, are one of the most effective ways to visualize data distribution and compare multiple datasets (batches) in a single graph.

A boxplot is a graphical representation of a dataset's distribution using five-number summary statistics: min, Q1, Q2, Q3 and max.

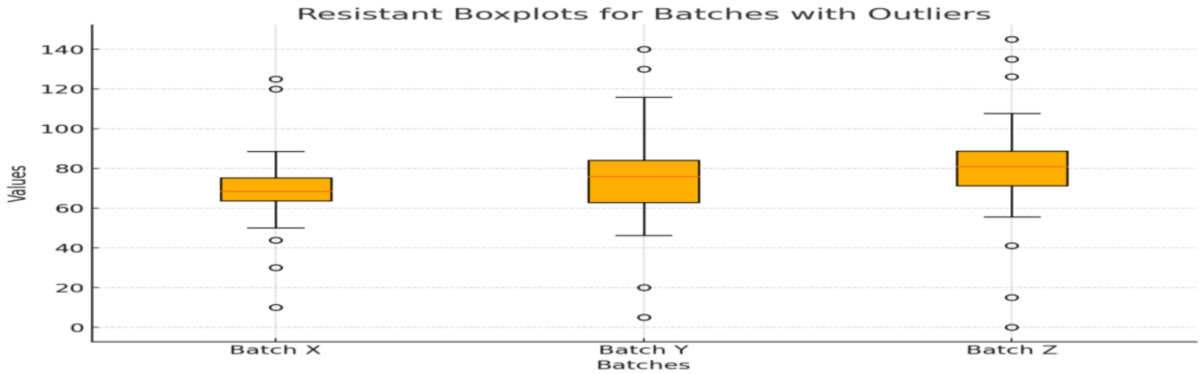
We can capture:

- location
- spread
- skewness
- tail length
- outlying data points

This compact visual display is especially useful for comparing several batches of data

Visualizing Boxplots with Batch Comparisons

Let's generate boxplots for three different batches of data and analyze their distributions (same n for each class).



In order to construct a resistant box plot, we need the 5-number summary; median, upper and lower fourth, cutoffs for outliers based on the fourth-spread.

How do you flag outliers?

Comparison batches using boxplot

A display of parallel boxplots can facilitate the comparison of several batches of data. From the display, we can see similarities and differences among the batches with respect to each of the five features already discussed.

Robust Measures of Spread and Variability

Classical measures of spread:

- Variance and Standard deviation: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Interquartile Range (IQR): IQR = Q3-Q1

Other alternatives:

- MAD (Median Absolute Deviation): A robust alternative to standard deviation.

$$\text{MAD} = \text{median}\ (|x_i - \text{median}(x)|)$$

Example:

Suppose we are given the following dataset:

$$X = \{2, 3, 5, 7, 10, 12, 15, 18, 50\}$$

Mean Calculation: $\bar{x} = 13.56$

Standard deviation:

$$s = \sqrt{\frac{1624.69}{8}} \approx 14.30$$

Computing MAD:

Median $(X) = 10$

Median Absolute Deviations, MAD = 5

Comment!

Influence function (IF): sensitivity of an estimator to small changes

The influence function (IF) of an estimator quantifies how sensitive it is to the small amount of contamination introduced into the data.

Influence Functions for common Estimators: Derivation

Mathematically, the IF of an estimator T under a distribution F is:

$$\mathrm{IF}(x; T, F) = \lim_{\epsilon \to 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon}$$

where:

- F is the true data distribution
- F_ε is the contaminated distribution, where an ε fraction of the data has been replased with x.
- $T(F)$ is the estimator applied to the original data.
- $T(F_\varepsilon)$ is the estimator applied to the contaminated data.

Interpretation:

- If the IF is unbounded the estimator is highly sensitive to small changes.
- If the IF is bounded, the estimator is robust to outliers.

(More of the details from the mathematical approach will be discussed later, during the course).

Some exercise!

We start with a clean dataset: $X = \{10, 12, 14, 16, 18, 20, 22, 24, 26, 28\}$

- Mean: $\bar{x} = 19$
- Median: $\tilde{x} = 19$

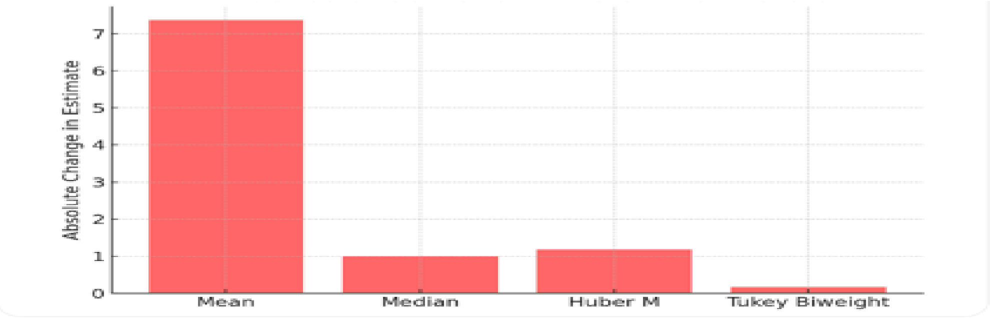
Effect of an outlier at $x = 100$

Let's introduce one outlier (100) into the dataset and compare its effect.

Estimator	Original value	New value (with outlier)	Change
Mean $\bar{x}$	19.0	27.1	+8.1 (unbounded)
Median $\tilde{x}$	19.0	19.0	0.0
Huber M-estimator (($c = 1.5$)	19.0	-20.5	+1.5 (moderate)
Tukey's Biweight	19.0	19.0	0.0 (ignore outlier)

- Mean is heavily affected (unbounded IF)
- Median remains unchanged (bounded IF)
- Huber M-estimator (slight shifts, downweights outlier)
- Tukey's Biweight completely ignores the outlier.

Influence of outlier(s) on different estimators: visualizing the IF behaviour



The bar chart illustrates the influence of an outlier on different estimators:

- Mean shows the largest change, confirming its unbounded IF.
- Median remains unchanged, demonstrating its bounded IF and robustness.
- Huber M-estimator shifts slightly, showing moderate resistance to the outlier.
- Tukey's Biweight estimator completely ignores the outlier, confirming its redescending IF.

Note:

1. IF quantifies an estimator's sensitivity to small data changes.
2. Bounded IF improves robustness by reducing the impact of outliers.
3. Redescending estimators (like Tukey's Biweight) completely ignore the outlier, confirming its redescending IF.

The Breakdown Point

The breakdown point is, roughly, the largest amount of contamination that may not cause the estimator to take on an arbitrary value.

Formally from Hampel & Stahel, 1982 and Donoho & Huber, 1983:

In a given sample $x_1, \cdots, x_n$, replace m data points $x_{i_1}, x_{i_2}, \cdots, x_{i_m}$ by arbitrary values $y_1, y_2, \cdots, y_m$.

Call the new data set $z_1, z_2, \cdots, z_n$.

The (lower finite sample gross error) breakdown point is

$$\varepsilon_n^*(T; x_1, x_2, \cdots, x_n) = \max\left\{\frac{m}{n}; \max_{i_1, \cdots i_m} \sup_{y_1, \cdots, y_m} |T(z_1, z_2, \cdots, z_n)| < \infty\right\}$$

Typically, ε_n^* is independent of $x_1, \cdots, x_n$.

Examples of breakdown points

Estimator	Breakdown point
Mean	0%
Median	50%
Trimmed mean (20%)	20%
Huber M-estimator	Approx. 30%
Least squares regression	0%
Least median of squares	50%

Key insight:

- The higher the breakdown point, the more robust the estimator.
- The median (50% breakdown point) is the most robust location estimator.
- The mean (0% breakdown point) is highly sensitive to even a single extreme outlier.

Some Practical Applications:

- In anomaly detection cybersecurity, a compromise estimator like Huber M-estimator is often used to detect fraudulent behaviour while minimizing false positives.
- Economics & finance: trimmed means are used to estimate inflation rates by removing extreme price changes.
- Medical statistics: median survival time in clinical trials is often preferred over the mean, which can be skewed by extreme survival times.
- Image Processing: In computer vison, robust estimators (e.g. Tukey's biweight function) are used to filter noise from, images, ensuring that a few extreme pixel values do not distort the entire image.

Final comments:

We want:

- Bounded influence function (with low gross-error sensitivity);
- High breakdown point;
- Consistensy (Fisher's)

How do we choose a good (robust) estimator?

Often look for estimators with the above property that are also highly efficient (optimal?) at an assumed central model. Compromises must be made to achieve good overall performance.

Topics include (as time permits):

1. Introduction. Examples Basic concepts, equivariance, breaking point,

2. Location/scale estimates, M-estimates, Pseudo-observations, depth

3. L-statistics: Linear combination of order statistics

4. Measures of robustness: Sensitivity curve, Influence Function

5. Numerical computation of M-estimates, Iterative Reweighted Least Square (IRLS)

6. Smoothing non smooth problems

7. Robust regression for multivariate statistics

8. Quantile regression