Lecture 1: Descriptive Statistics

and Empirical Distributions

You get a spreadsheet with 10,000 rows. . .

What do you look at first?

And what can fool you?

Before any model, any test, any estimation — **look at the data**.

Goals of Descriptive Statistics

Summarize **center**, **spread**, and **shape**

Detect **outliers**, missing data, impossible values

Compare groups or time periods visually

Generate **hypotheses** before testing them

Descriptive $\neq$ inferential: we're describing *this sample*, not yet the population.

Measures of Center

Sample Mean

$\bar{X} = \frac{1}{n}\sum X_i$

Uses all data
Minimizes squared error

Sensitive to outliers

Sample Median

Middle value

Robust (50% breakdown)
Ignores magnitudes

Resists outliers

Mode

Most frequent value

Best for categorical
Can be non-unique

Minimizes 0–1 loss

Trimmed Mean

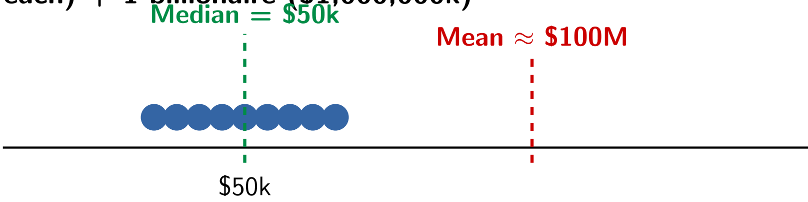
Drop top/bottom $k\%$

Compromise: mean $\leftrightarrow$ median

Tunable robustness

The Billionaire in the Room

9 teachers (salary \$50k each) + 1 billionaire (\$1,000,000k)



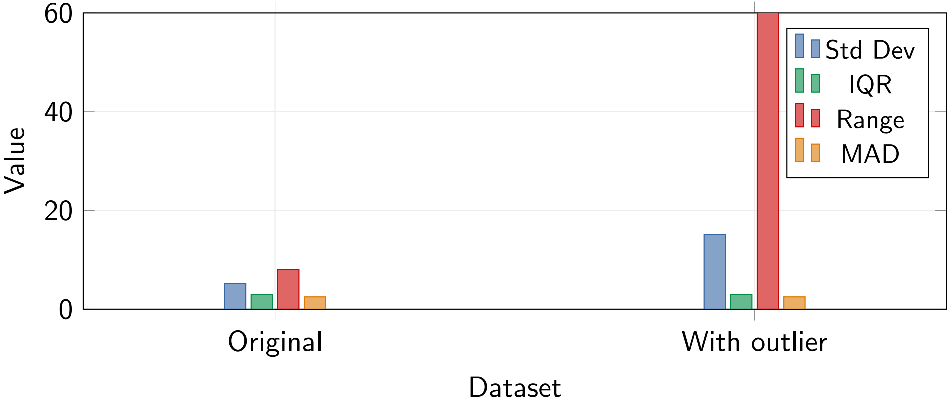
Which better describes a "typical" person in this roon

Measures of Spread

Measure	Formula	Properties
Variance S^2	$\frac{1}{n-1}\sum(X_i - \bar{X})^2$	Uses all data; $n-1$ = Bessel's correction (unbiased)
Std Dev S	$\sqrt{S^2}$	Same units as data
Range	$\max - \min$	Simple; extremely fragile
IQR	$Q_3 - Q_1$	Middle 50%; robust
MAD	$\text{med}\,\lvert X_i - \text{med}\rvert$	Most robust; companion to median
CV	$S/\bar{X}$	Dimensionless; compare across scales

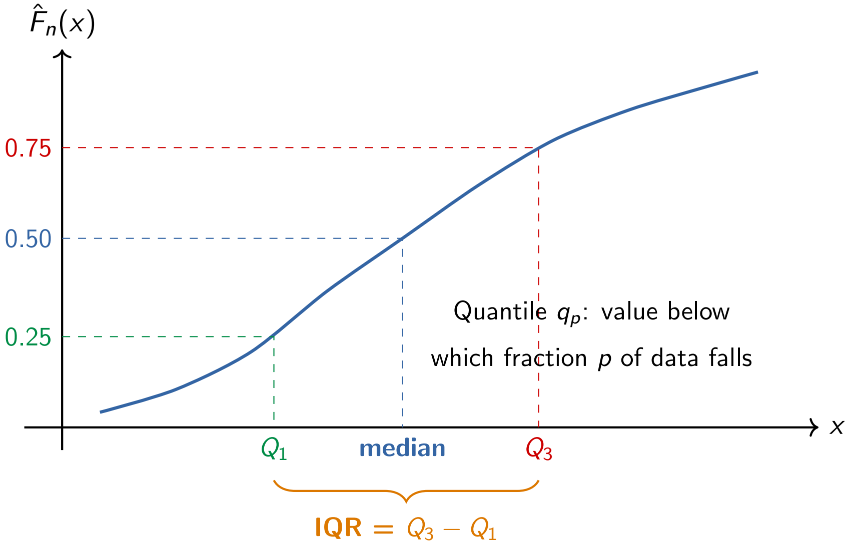
Why $n-1$? We used up one "degree of freedom" estimating $\bar{X}$.
(We'll prove $\mathbb{E}[S^2] = \sigma^2$ in Lecture 3.)

Robust vs Non-Robust: Visual

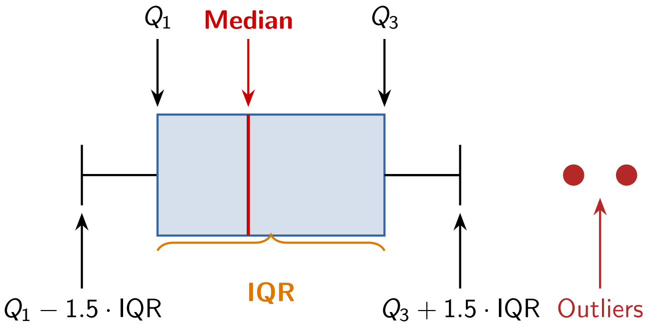


One outlier: Range explodes, Std Dev triples. IQR and MAD don't budge.

Quantiles and Percentiles



Boxplot Anatomy



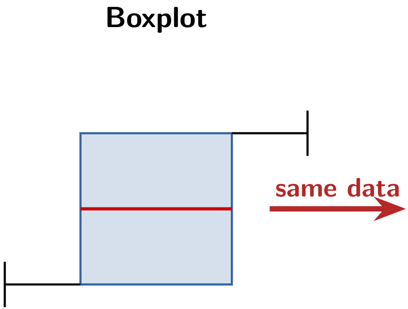
Strengths:

- Compact group comparison
- Shows center, spread, outliers

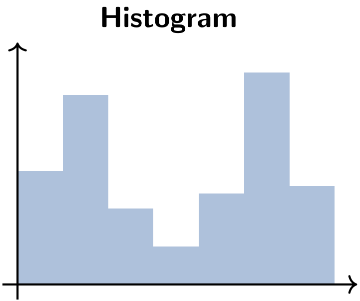
Weakness:

- Hides multimodality!
- Pair with histogram or violin

Boxplot Hides Bimodality



Looks unimodal



Two distinct groups!

Always pair boxplots with histograms or violin plots.

Quantiles in the Real World

Finance: Value at Risk

VaR = 5th percentile of the loss distribution
"Worst 5% scenario"

Medicine: Growth Charts

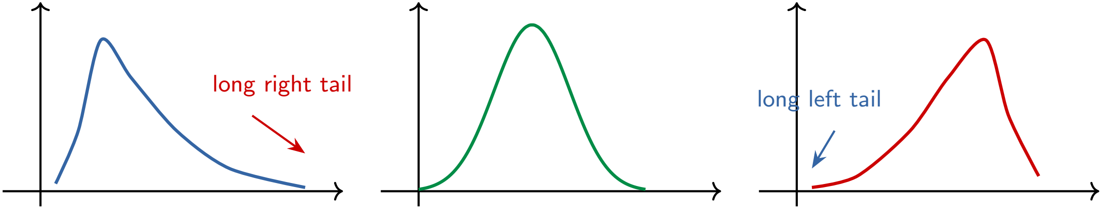
Child's height at the 3rd, 50th, 97th percentile relative to age group

Education: Test Scores

"You scored in the 85th percentile" = better than 85% of test-takers

Skewness: Measuring Asymmetry

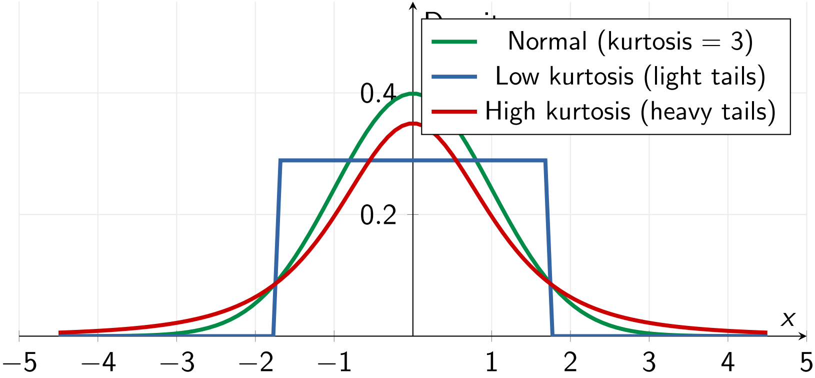
Positive Skew | **Symmetric** | **Negative Skew**



Income, house prices | Heights, measurement error | Exam scores near ceiling

$$\text{Skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S}\right)^3$$

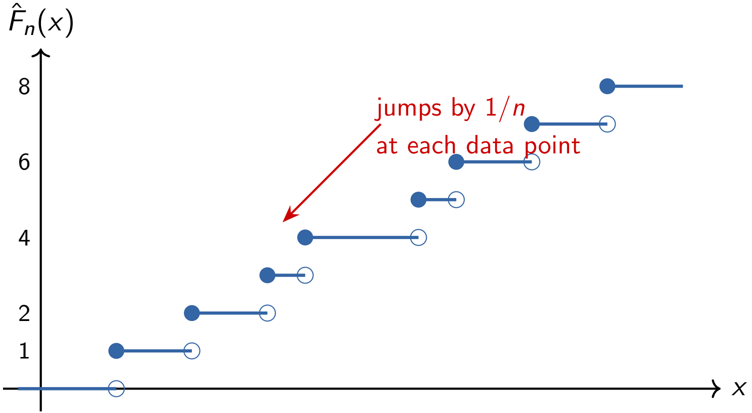
Kurtosis: Tail Heaviness



High kurtosis $\Rightarrow$ more extreme outliers than normal predicts.
Financial returns have high kurtosis — assuming normality underestimates risk.

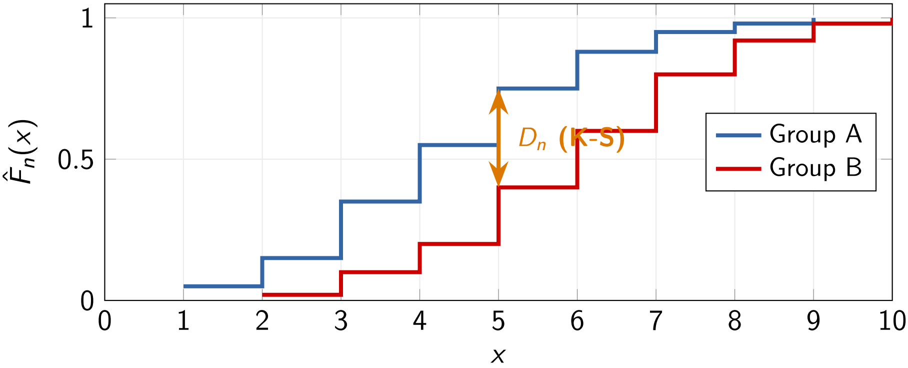
The Empirical CDF

$$\hat{F}_n(t) = \frac{1}{n}\#\{X_i \leq t\} = \frac{\text{number of observations} \leq t}{n}$$



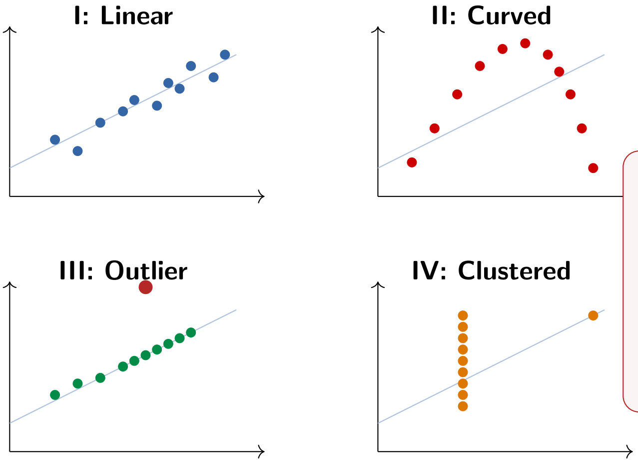
Example: $n = 8$ observations

ECDF: Why It's Powerful



- No bin-width choice (unlike histograms) — the ECDF is **parameter-free**
- Biggest gap = **Kolmogorov–Smirnov statistic** (formalized in Lecture 7)
- Glivenko–Cantelli: $\hat{F}_n \to F$ uniformly as $n \to \infty$

Anscombe's Quartet (1973)



All four datasets:

$$\begin{aligned} \bar{x} &= 9,\ \bar{y} \approx 7.5 \\ S_x^2 &= 11,\ S_y^2 \approx 4.13 \\ r &\approx 0.816 \\ \hat{y} &= 3 + 0.5x \end{aligned}$$

Identical statistics.
Wildly different data.

Simpson's Paradox

Overall:
Men admitted at higher rate
$\Rightarrow$ "Gender bias?"

Dept A (easy):
Women: 82%
Men: 62%

Dept B (hard):
Women: 68%
Men: 63%

Dept C (hard):
Women: 34%
Men: 33%

Women applied to **more competitive** departments.
Within each department, women were admitted at **equal or higher** rates.

Aggregate trend reversed inside subgroups!

Practical: Exploratory Data Analysis

Pick a real dataset (Titanic, Palmer Penguins, or your own):

1. Compute mean, median, SD, IQR, skewness for each numeric variable
2. Make **histograms** — identify skewed variables, outliers, multimodal shapes
3. Make **boxplots by group** (e.g., survival by class, mass by species)
4. Plot **ECDFs** for two subgroups on the same axes
5. Find a case where a summary statistic is **misleading** and a plot reveals the truth

Bonus: Construct your own Anscombe-style pair — two tiny datasets with the same mean and variance but different shapes.

Questions?

Next lecture: Point Estimation — Maximum Likelihood