Lecture 2: Descriptive Statistics

Center · Spread · Quantiles · Shape · ECDF · Anscombe · Simpson

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

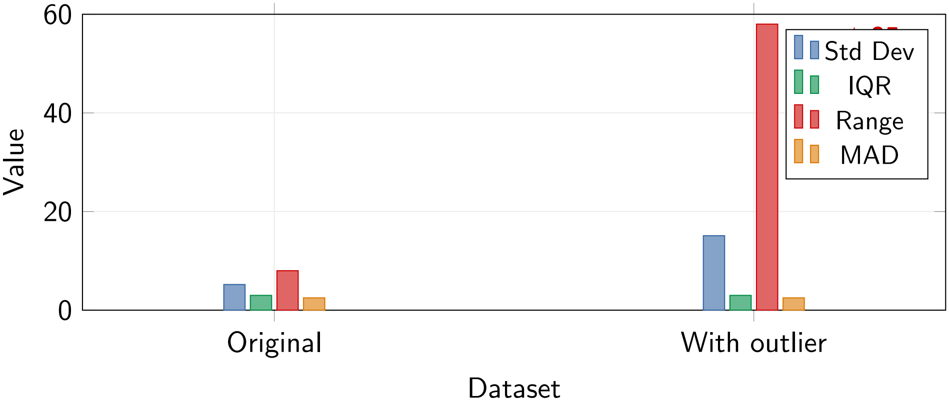
Sample Mean	Sample Median	Mode	Trimmed Mean
$\bar{X} = \frac{1}{n}\sum X_i$	Middle value	Most frequent value	Drop top/bottom $k\%$
Uses all data Minimizes squared error	Robust (50% breakdown) Ignores magnitudes	Best for categorical Can be non-unique	Compromise: mean $\leftrightarrow$ median
Sensitive to outliers	Resists outliers	Minimizes 0–1 loss	Tunable robustness

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

Why $n-1$? We used up one "degree of freedom" estimating $\bar{X}$. This makes S^2 **unbiased**: $\mathbb{E}[S^2] = \sigma^2$.

Robust vs Non-Robust: Visual



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

How a Histogram Is Built

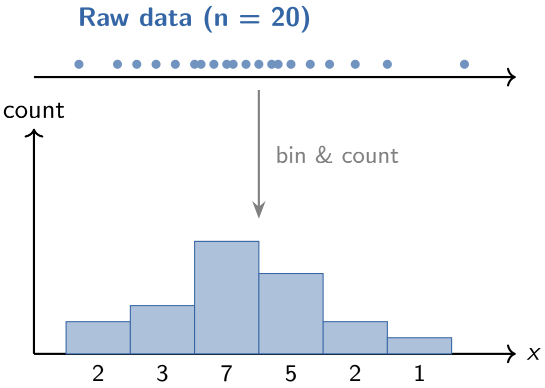
Recipe:

1. Choose **bins**: equal-width intervals covering the data range
2. Count observations in each bin
3. Draw bars — height = count (or density)

Density form:

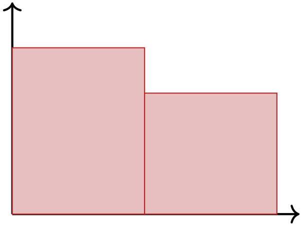
$$\text{height} = \frac{\text{count}}{n \times \text{bin width}}$$

so total area = 1 (comparable across bin widths).



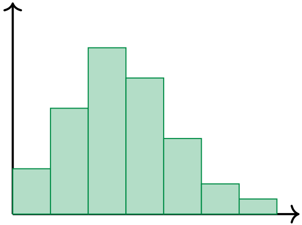
Bin Width Matters

Too few bins (2)



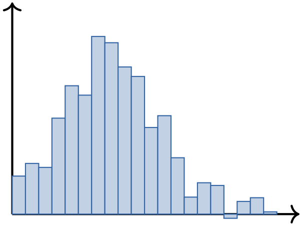
Hides all structure

Good bin width



Shape is clear

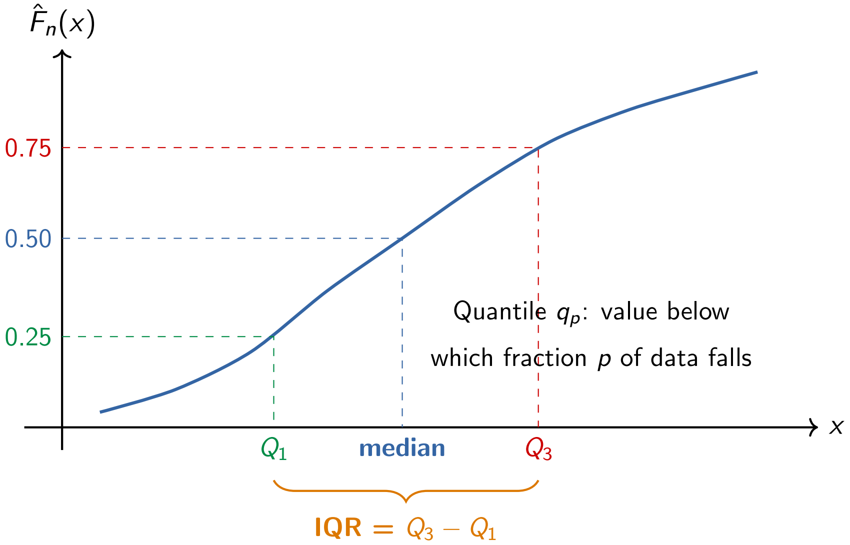
Too many bins (20)



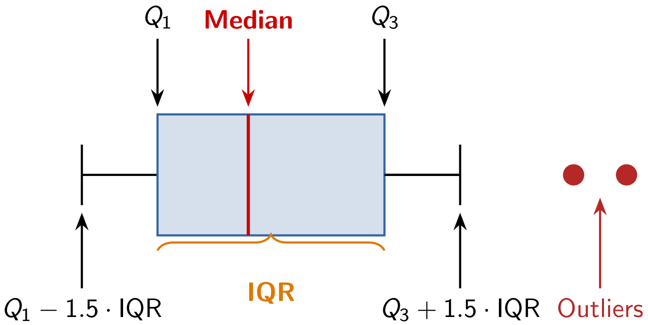
Too noisy

Common rules: Sturges ($k = 1 + \log_2 n$), Freedman–Diaconis ($h = 2 \cdot \mathrm{IQR} \cdot n^{-1/3}$).
In practice: try several and look.

Quantiles and Percentiles



Boxplot Anatomy



Five-number summary: min, Q_1, median, Q_3, max — the boxplot visualizes exactly this.

Strengths:

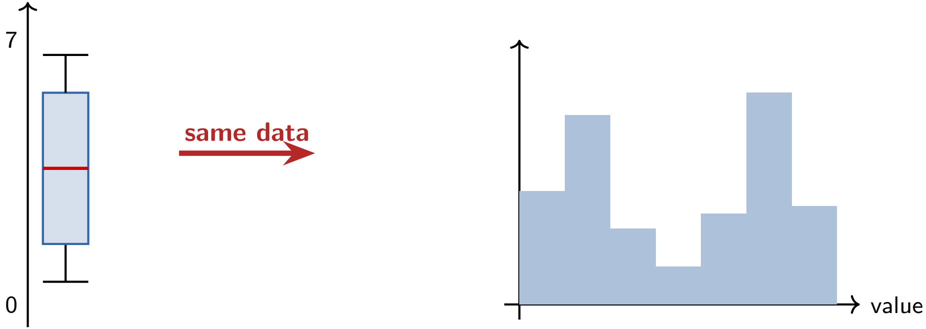
- Compact group comparison
- Shows center, spread, outliers

Weakness:

- Hides multimodality!
- Pair with histogram or violin

Boxplot Hides Bimodality

Boxplot **Histogram**

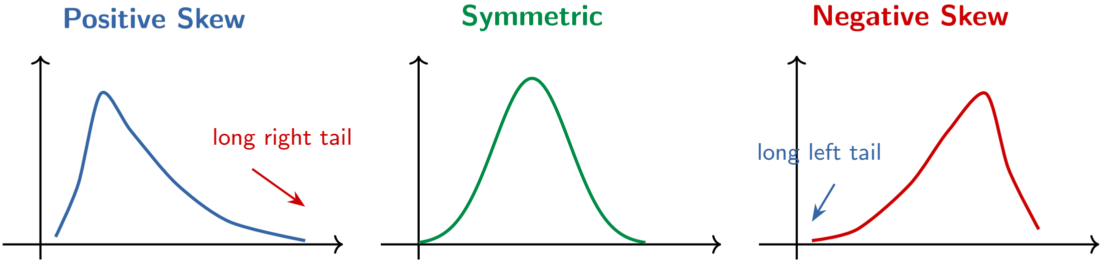


Looks unimodal

Two distinct groups!

Always pair boxplots with histograms or violin plots.

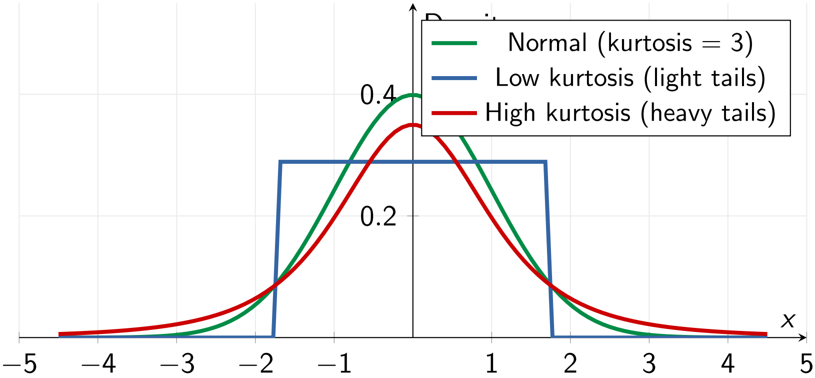
Skewness: Measuring Asymmetry



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

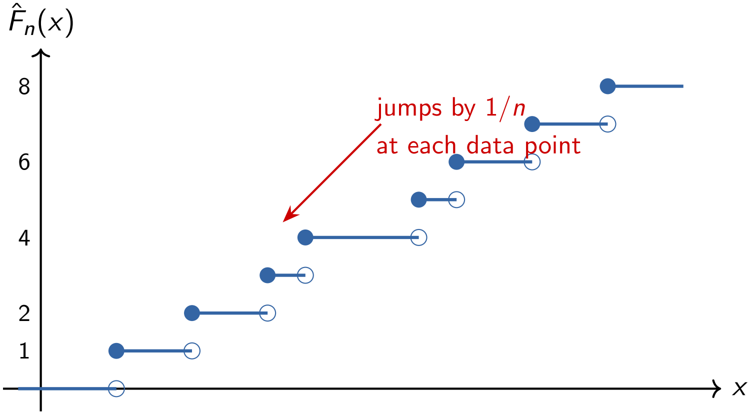


$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S}\right)^4 \qquad \text{Excess kurtosis} = \text{Kurt} - 3$$

Normal has kurtosis = 3 (excess = 0). Most software reports **excess kurtosis**.
Financial returns have high kurtosis — assuming normality underestimates risk.

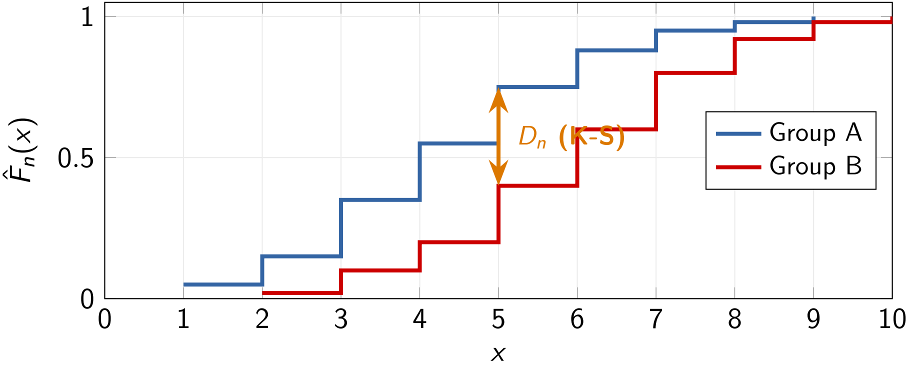
The Empirical CDF

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



Example: $n = 8$ observations

ECDF: Why It's Powerful



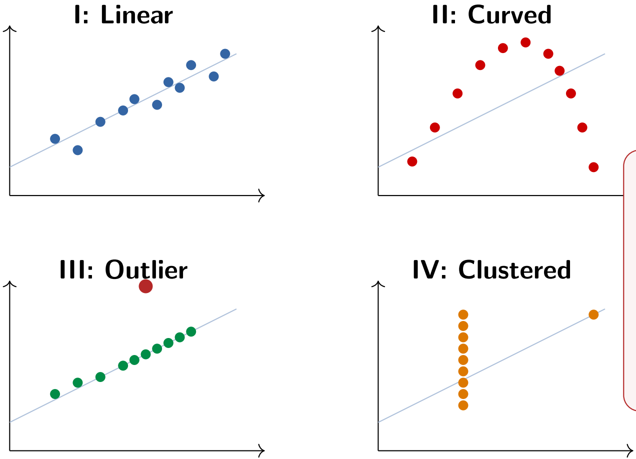
- No bin-width choice (unlike histograms) — the ECDF is **parameter-free**
- Biggest gap = **Kolmogorov–Smirnov statistic** D_n
- Glivenko–Cantelli: $\hat{F}_n \to F$ uniformly as $n \to \infty$

Choosing the Right Plot

What do I want to see?	Best plot
Distribution shape of one variable	**Histogram** / density plot
Compare distributions across groups	**Boxplot** / violin plot
Relationship between two variables	**Scatter plot**
Full distribution, no bin choices	**ECDF**
Change over time	**Line plot** / time series

Rule of thumb: always start with a histogram + scatter plot matrix. Then refine.

Anscombe's Quartet (1973)

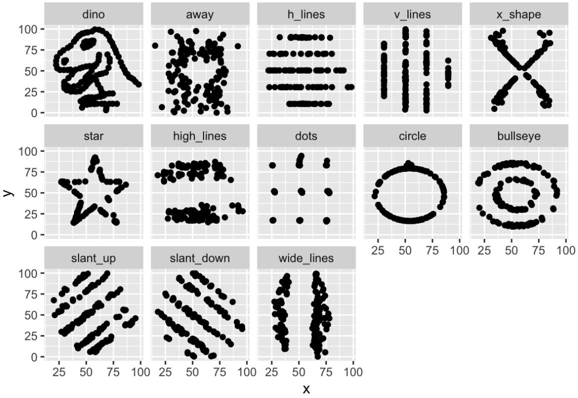


All four datasets:

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

Identical statistics.
Wildly different data.

The Datasaurus Dozen (2017)



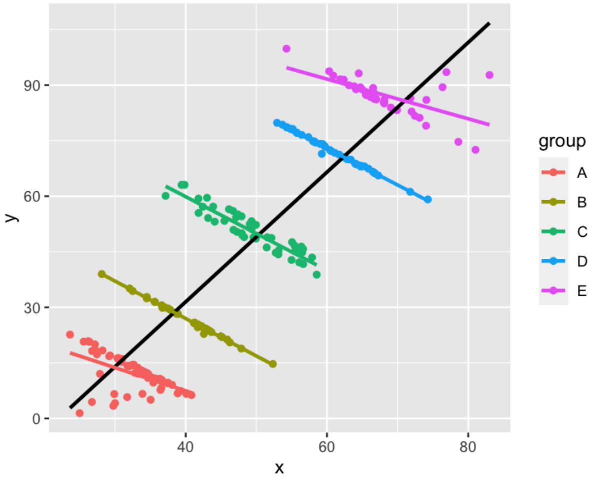
13 datasets, all with:

- Same $\bar{x}$, $\bar{y}$
- Same S_x, S_y
- Same correlation r

Yet shapes include a **dinosaur**, a star, parallel lines, a circle. . .

Never trust summary statistics alone. Always plot your data.

Simpson's Paradox



Each colored group trends **down**, yet the aggregate trend goes **up**.
How? The groups have **different sizes and positions**.

Simpson's Paradox: UC Berkeley Admissions (1973)

12,763 applicants to UC Berkeley graduate programs.

Aggregate data:

	Applied	**Admitted**
Men	8,442	44%
Women	4,321	35%

9 percentage points gap!
Lawsuit filed for gender bias.

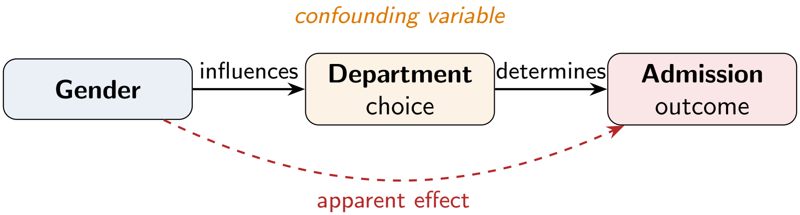
By department:

	Rate	**Women**	**Men**
Dept A	easy	82%	62%
Dept B	easy	68%	63%
Dept C	hard	34%	35%
Dept D	hard	7%	6%

Women admitted at equal or **higher** rates in each dept!

How is this possible?

Simpson's Paradox: Why It Happens



The key: Women disproportionately applied to **hard** departments (low acceptance for everyone). Men disproportionately applied to **easy** departments.

When you aggregate, the **different weights** reverse the trend:

- Women: ~80% applied to hard depts $\rightarrow$ low overall rate
- Men: ~80% applied to easy depts $\rightarrow$ high overall rate

General lesson: a trend in every subgroup can **reverse** when subgroups are combined.
Always ask: *is there a hidden variable that changes the group sizes?*

Questions?