Descriptive Statistics

Lesson 02

Digital Finance

Outline

What is Descriptive Statistics?

Descriptive statistics summarize and describe data

- Organize raw data into meaningful information
- Identify patterns, trends, and outliers
- Communicate findings clearly

Two main categories:

- **Numerical summaries**: Mean, variance, percentiles
- **Visual summaries**: Histograms, boxplots, scatter plots

Descriptive statistics tell us what the data looks like.

The Mean (Average)

Intuition: Add up all values, divide by how many there are.
Example: Test scores 70, 80, 90 $\rightarrow$ Mean $= (70 + 80 + 90)/3 = 80$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Key properties:

- Uses all data points
- Sensitive to extreme values (one outlier can shift the mean)
- The "balance point" – data balances around the mean

The Σ symbol means "sum of" – we'll use it throughout this course.

The Median

Median: The middle value when data is sorted

- If n is odd: middle value
- If n is even: average of two middle values

Properties:

- Not affected by extreme values (robust)
- Always exists and is unique
- 50th percentile of the data

When to use:

- Skewed distributions (income, house prices)
- Data with outliers

The median splits the data in half.

The Mode

Mode: The most frequently occurring value

- Can have no mode, one mode, or multiple modes
- Useful for categorical data

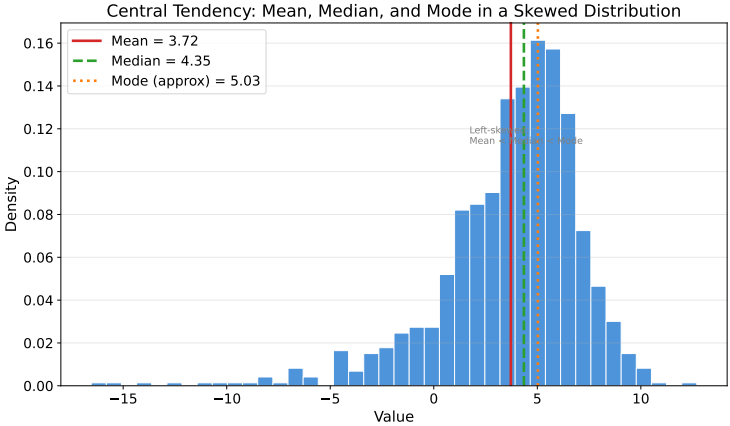
Types:

- **Unimodal**: One peak
- **Bimodal**: Two peaks
- **Multimodal**: Multiple peaks

Finance example: Most common credit rating among bonds

The mode tells us what value occurs most often.

Comparing Mean, Median, and Mode



In skewed data, mean is pulled toward the tail; median is more robust.

Which Measure to Use?

Situation	Best Measure	Reason
Symmetric data	Mean	Uses all information
Skewed data	Median	Robust to outliers
Extreme outliers	Median	Not affected
Categorical data	Mode	Only option
Income/wealth	Median	Right-skewed
Stock returns	Mean	Often symmetric

Choose the measure that best represents "typical" for your data.

Why Measure Spread?

Central tendency alone is insufficient

- Two datasets can have the same mean but very different spreads
- Spread measures variability, dispersion, or uncertainty

In finance, spread = risk:

- Higher spread in returns = higher volatility = higher risk
- Investors are compensated for taking risk

"Risk" in finance is often measured by spread.

Range and Interquartile Range

Range: Maximum minus minimum (simple but sensitive to outliers)
What are quartiles? Values that divide sorted data into 4 equal parts:

- Q_1 (25th percentile): 25% of data below this point
- Q_2 (50th percentile): The median
- Q_3 (75th percentile): 75% of data below this point

Interquartile Range (IQR): $Q_3 - Q_1$

- Contains the middle 50% of data
- Not affected by outliers (robust)

Percentile = percent of data below that value. Quartiles are the 25th, 50th, 75th percentiles.

Variance

Intuition: How far are data points from the mean, on average?
Steps: (1) Find distance from mean, (2) Square it, (3) Average

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

Why square the distances?

- Makes all distances positive (no canceling)
- Penalizes large deviations more than small ones

Why divide by $n - 1$? Technical correction for samples (not n).

Variance answers: "How spread out is the data?" Units are squared.

Standard Deviation

Standard Deviation: Square root of variance

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Advantages over variance:

- Same units as the original data
- More intuitive interpretation

In finance:

- Standard deviation of returns = **volatility**
- Key input for risk management and portfolio theory

Standard deviation is the most common measure of risk in finance.

Coefficient of Variation

CV: Standard deviation relative to the mean

$$CV = \frac{s}{\bar{x}} \times 100\%$$

When to use:

- Comparing variability across different scales
- Data with different units
- Relative (not absolute) variation matters

Example:

- Stock A: mean return 10%, std dev 15% $\rightarrow$ CV = 150%
- Stock B: mean return 5%, std dev 10% $\rightarrow$ CV = 200%

CV allows comparison of variability regardless of scale.

Why Visualize Data?

Pictures reveal what numbers cannot

- Patterns and trends
- Outliers and anomalies
- Distribution shape

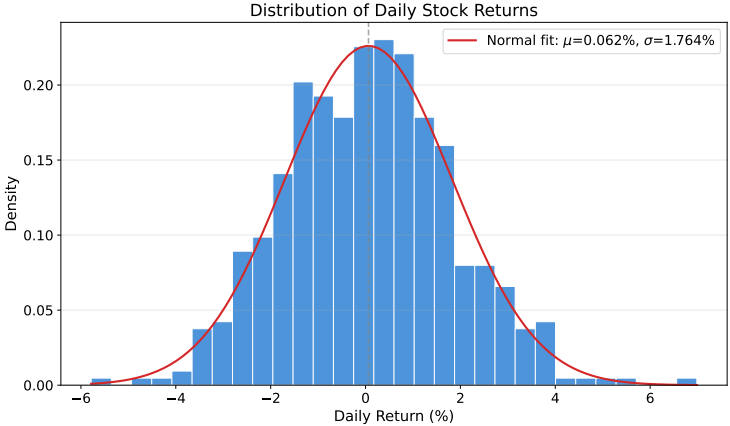
Common visualization types:

- **Histogram**: Distribution of one variable
- **Boxplot**: Summary statistics and outliers
- **Scatter plot**: Relationship between two variables

"A picture is worth a thousand numbers." – Francis Anscombe (attributed)

Histogram

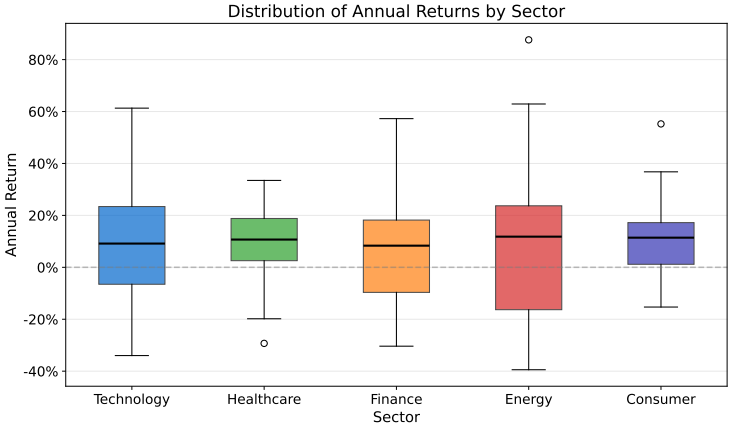
Shows the distribution of a single variable



The histogram reveals the shape, center, and spread of the distribution.

Boxplot (Box-and-Whisker Plot)

Summarizes five-number summary + outliers



Boxplots are excellent for comparing distributions across groups.

Reading a Boxplot

Components of a boxplot:

- **Box**: From Q_1 to Q_3 (contains middle 50%)
- **Line inside**: Median (Q_2)
- **Whiskers**: Extend to min/max (within 1.5 IQR)
- **Points**: Outliers (beyond 1.5 IQR)

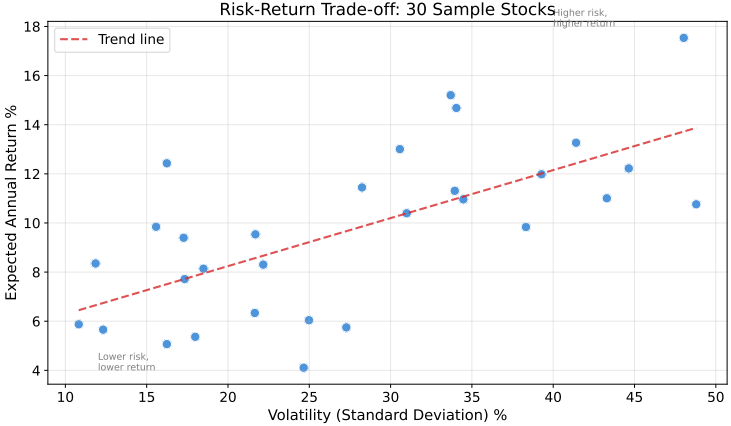
What to look for:

- Center (median position)
- Spread (box size)
- Skewness (asymmetry)
- Outliers (individual points)

A single boxplot shows five-number summary at a glance.

Scatter Plot

Shows relationship between two variables



The risk-return trade-off: higher volatility often means higher expected returns.

Skewness

Skewness: Which tail is longer? (Is the distribution lopsided?)

- **Symmetric**: Balanced left and right (skewness $= 0$)
- **Right-skewed**: Long right tail, most values on left (skewness > 0)
- **Left-skewed**: Long left tail, most values on right (skewness < 0)

Examples:

- Income: Right-skewed (few very high earners)
- Exam scores: Often left-skewed (ceiling effect)
- Stock returns: Slightly left-skewed (crashes ¿ booms)

Formula exists but focus on interpretation. See next slide for visual examples.

Kurtosis

Kurtosis: How heavy are the tails? (How common are extreme values?)

- **Normal tails**: Like a normal distribution (baseline)
- **Fat tails**: More extreme values than normal (common in finance!)
- **Thin tails**: Fewer extreme values than normal

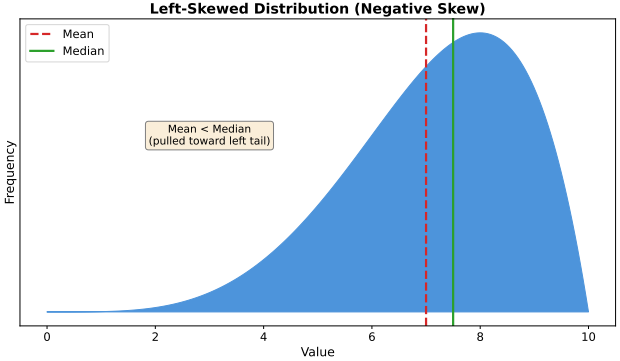
Excess kurtosis = Kurtosis − 3 (so normal = 0)

- Positive excess kurtosis = fat tails = more surprises
- Negative excess kurtosis = thin tails = fewer surprises

Why it matters: Stock returns have fat tails – crashes and booms happen more often than a normal distribution predicts!

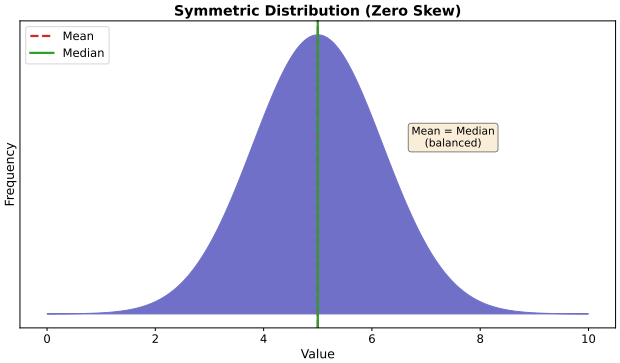
Formula is complex (see textbook). Focus on interpretation: fat tails = more extreme events.

Left-Skewed Distribution



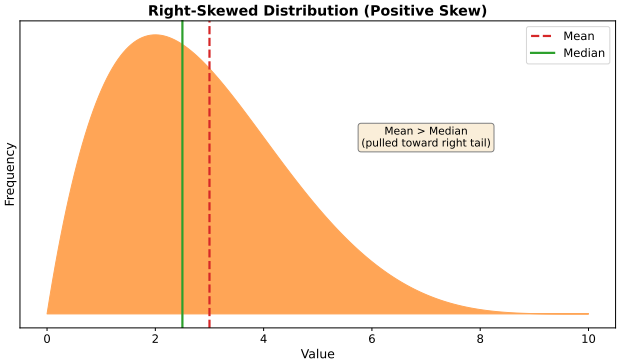
In left-skewed data, the mean is pulled toward the long left tail.

Symmetric Distribution



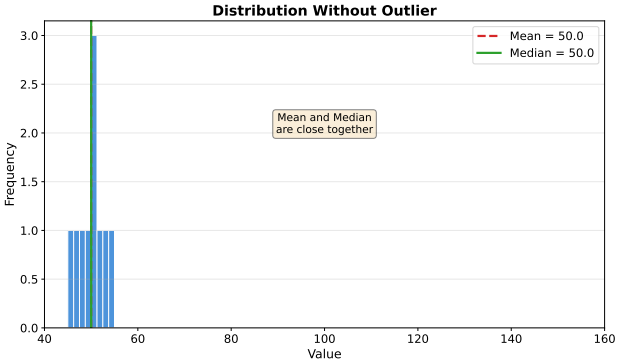
In symmetric data, mean and median are equal.

Right-Skewed Distribution



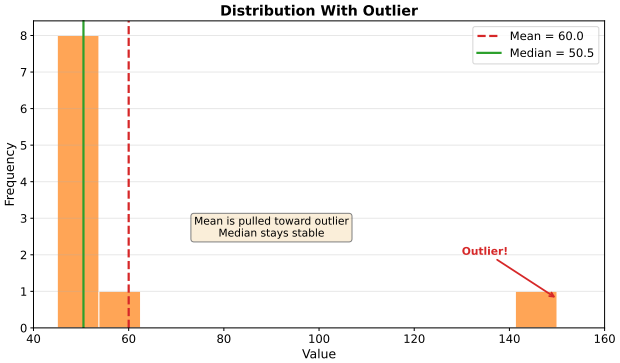
In right-skewed data (like income), the mean is pulled toward the long right tail.

Without Outliers



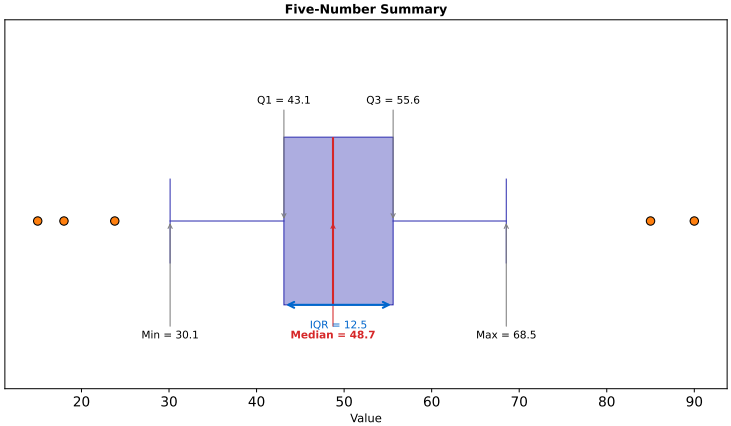
Without outliers, mean and median give similar results.

Impact of Outliers



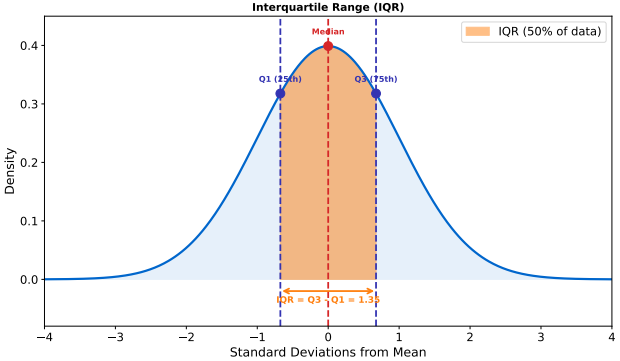
One outlier shifts the mean dramatically, but the median stays stable.

Five-Number Summary



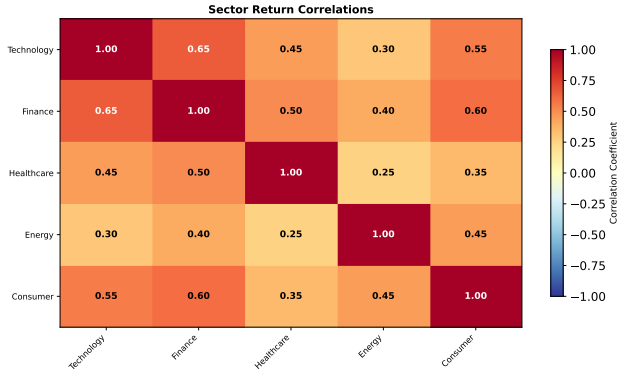
Min, Q1, Median, Q3, Max provide a complete picture.

Interquartile Range Visualization



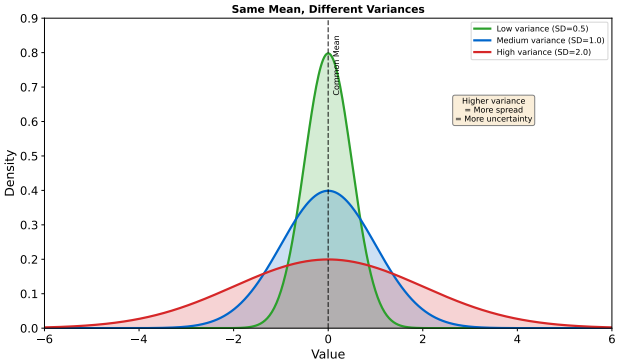
The IQR contains the middle 50% of the data.

Correlation Heatmap



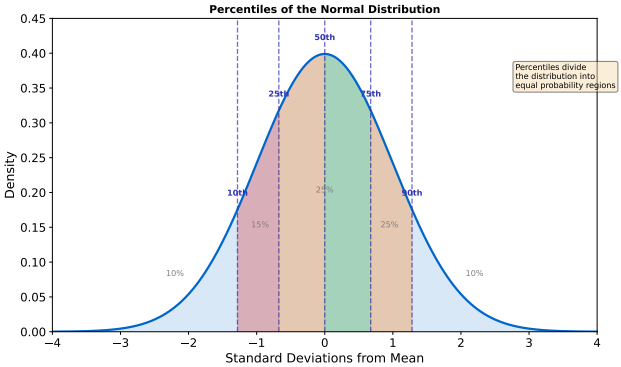
Correlation measures linear relationship: +1 = perfect positive, -1 = perfect negative, 0 = none. Details in Lesson 8.

Variance Comparison



Same mean, different variances illustrate the importance of spread.

Percentile Distribution



Percentiles divide the distribution into 100 equal parts.

Summary

Central tendency:

- Mean: average, sensitive to outliers
- Median: middle value, robust
- Mode: most frequent value

Spread:

- Variance and standard deviation: most common
- Range and IQR: simpler alternatives

Shape:

- Skewness: asymmetry
- Kurtosis: tail heaviness

Next lesson: Set Theory and Probability Axioms