

# Lesson 13: Descriptive Statistics

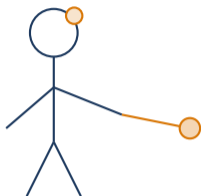
Data Science with Python – BSc Course

Data Science Program

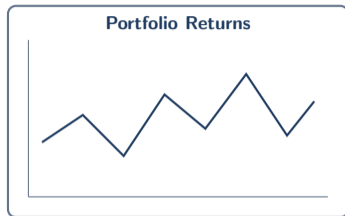
BSc Course

45 Minutes

# Your Portfolio's Vital Signs



Data Scientist



Vital Signs
Mean: 8.2%
Std Dev: 15.3%
Skew: -0.4
Kurtosis: 4.1
Sharpe: 0.54

**Statistics is a medical checkup for your data.**

Today we learn to read the vital signs of any dataset.

---

Let's check your portfolio's vital signs

# Learning Objectives

After this lesson, you will be able to:

1. **Explain** why central tendency matters for summarizing data
2. **Compare** mean vs median and when each is appropriate
3. **Calculate** dispersion measures (variance, std dev, IQR)
4. **Interpret** skewness and kurtosis in financial returns
5. **Apply** risk metrics (Sharpe, VaR, drawdown) to portfolios

---

Bloom's levels: explain, compare, calculate, interpret, apply

# From DataFrames to Statistical Insight

## What you already know:

- Python basics (L01–L06)
- DataFrames and selection (L05–L06)
- GroupBy and operations (L08–L09)

## What this lesson adds:

- Statistical summaries of data
- Shape of distributions
- Risk metrics used in finance

**You can load and manipulate data.  
Now learn to describe what it means.**

---

Statistics turns raw numbers into actionable insight

# The Central Challenge

How do you summarize 10,000 data points into insights you can act on?

2.1 -1.3 0.7 -0.2 1.5 -0.8 0.3 -1.1 ...



Mean = 0.15% Std = 1.8% Skew = -0.4  
"Slightly negative, volatile returns"

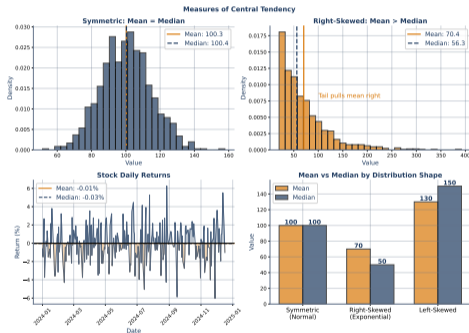
---

Descriptive statistics compress data into a handful of meaningful numbers

# Central Tendency: Where Is the Center?

Three measures of the “typical” value:

- **Mean** – average:  $\bar{x} = \frac{1}{n} \sum x_i$
- **Median** – middle when sorted
- **Mode** – most frequent

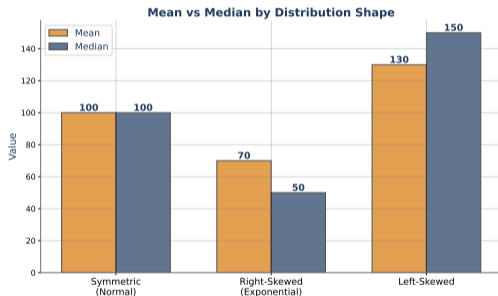


Mean uses every data point; median ignores magnitude; mode counts frequency

# When Mean $\neq$ Median

## Outliers pull the mean but not the median

- Symmetric data: mean  $\approx$  median
- Skewed data: mean shifts toward the tail



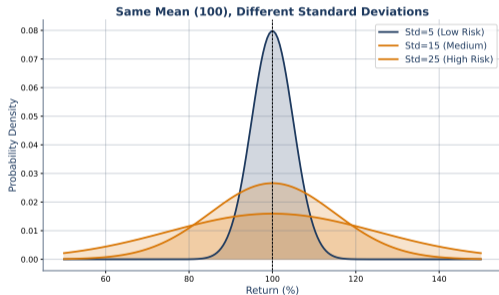
---

Median is robust to outliers; mean captures total magnitude

# Dispersion: How Spread Out Is the Data?

Variance and standard deviation quantify spread:

- Variance  $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$
- Std dev  $s = \sqrt{s^2}$  (same units as data)



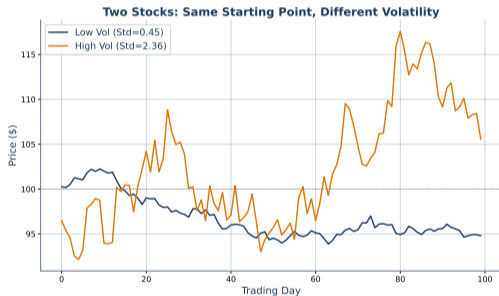
---

Same mean, different standard deviations – spread matters

# Std Dev as Financial Volatility

In finance, standard deviation = volatility

- Higher volatility = higher risk and potential reward
- Annualized:  $\sigma_{annual} = \sigma_{daily} \times \sqrt{252}$



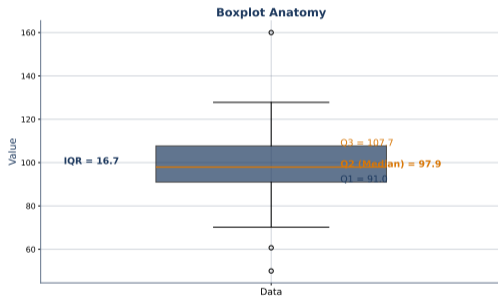
---

Volatility is the most common risk measure in portfolio management

# Quartiles: The Boxplot

## Five-number summary + outliers in one chart:

- Box: Q1 (25%) to Q3 (75%), line at median
- Whiskers:  $1.5 \times \text{IQR}$  beyond box; dots are outliers

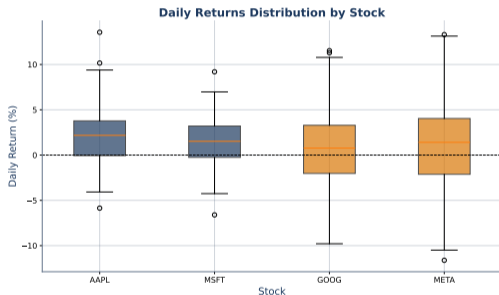


$\text{IQR} = \text{Q3} - \text{Q1}$ ; outliers lie beyond  $1.5 \times \text{IQR}$  from box edges

# Comparing Distributions Across Assets

Boxplots side-by-side reveal differences at a glance

- Median: which asset has higher typical return?
- Box width: which is more volatile?



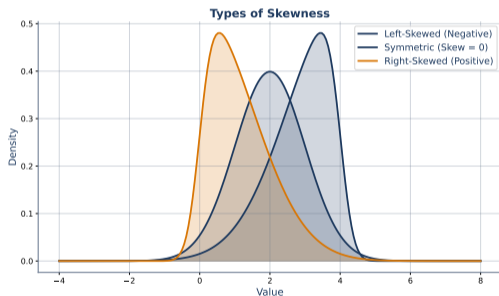
---

Compare return distributions across assets in one chart

# Skewness: Is the Distribution Symmetric?

Skewness measures asymmetry:

- Positive skew: long right tail (mean  $>$  median)
- Negative skew: long left tail (mean  $<$  median)



---

Stock returns often show negative skew – large losses are more likely than large gains

# Checkpoint: Skewness in Practice

## Think about it:

Is positive skewness **good** or **bad**  
for an investor?

Hint: Where is the long tail – gains or losses?

### Positive skew

Long right tail = rare  
large *gains* possible

### Negative skew

Long left tail = rare  
large *losses* possible

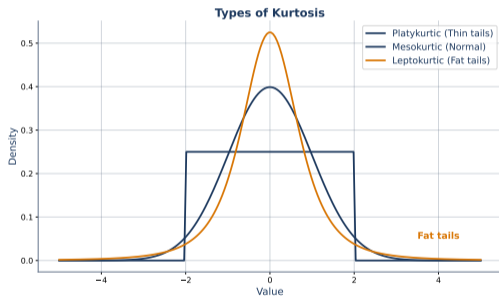
---

Investors prefer positive skew – upside surprises over downside crashes

# Kurtosis: How Heavy Are the Tails?

Kurtosis measures tail extremity:

- **Leptokurtic** (excess  $> 0$ ): fat tails, more extremes
- **Platykurtic** (excess  $< 0$ ): thin tails, fewer extremes



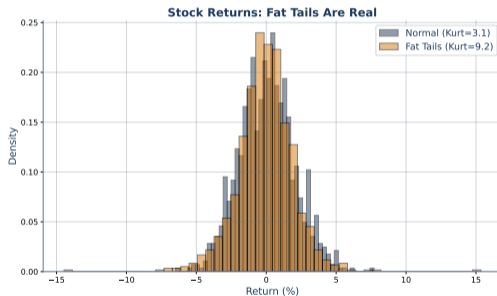
---

Pandas `df.kurtosis()` returns excess kurtosis (normal = 0)

# Fat Tails: Why They Matter in Finance

## Financial returns have fatter tails than normal:

- “Once in a century” events happen every decade
- Risk models that ignore fat tails fail catastrophically



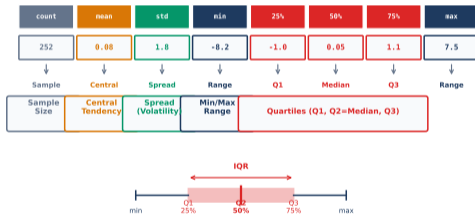
2008 crash: –20% days that normal models rated as “virtually impossible”

# One Command: `df.describe()`

Pandas gives you a statistical checkup in one line:

- Count, mean, std, min, Q1, median, Q3, max
- Add skew/kurtosis: `df.agg(['skew', 'kurtosis'])`

`df.describe()` Output Explained

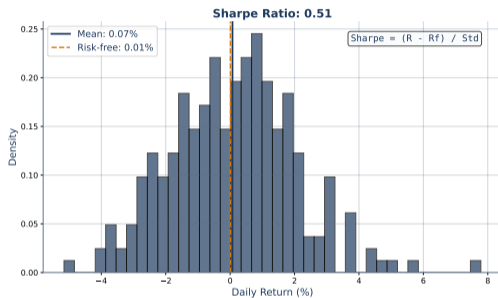


---

Think of `describe()` as the blood test results for your dataset

# Sharpe Ratio: Risk-Adjusted Return

Return per unit of risk:  $SR = \frac{\bar{r} - r_f}{\sigma_r}$     **Benchmarks:** >1.0 good, >2.0 excellent



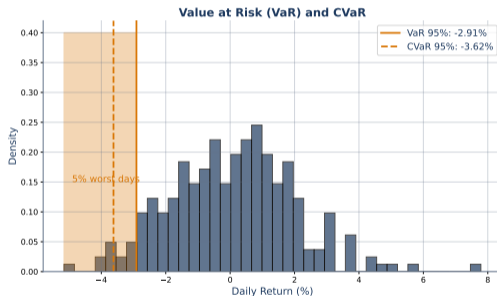
---

Sharpe ratio: the single most common performance metric in finance

# Value at Risk and Expected Shortfall

## VaR: “How bad can a bad day get?”

- VaR at 5%: loss exceeded only 5% of the time
- CVaR: average loss *when* VaR is breached

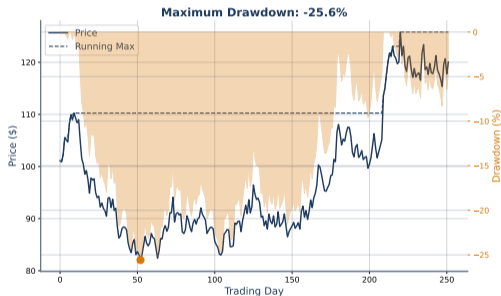


VaR answers “how bad?”; CVaR answers “how bad on average when bad?”

# Maximum Drawdown

Largest peak-to-trough decline:

- Worst cumulative loss an investor experiences
- Key metric for hedge funds and risk management

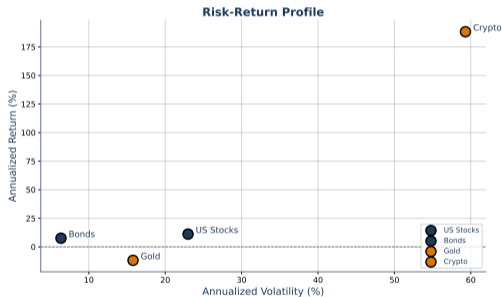


Max drawdown captures the pain of holding through a crash

# Risk vs Return: The Fundamental Tradeoff

Every asset lives on the risk-return plane:

- $x$ : volatility (risk),  $y$ : mean return (reward)
- Upper-left is ideal: high return, low risk



---

Plot return vs volatility to compare assets at a glance

# Hands-on Exercise (25 min)

## Analyze a stock portfolio with descriptive statistics:

1. Load daily returns for 4 stocks using `pd.read_csv()`
2. Run `df.describe()` and interpret each statistic
3. Calculate skewness and kurtosis – which stock has the fattest tails?
4. Compute Sharpe ratio for each stock ( $r_f = 0$ )
5. Create a risk-return scatter plot (volatility vs mean return)

**Bonus:** Calculate 5% VaR and CVaR for each stock. Which has the highest tail risk?

---

Apply everything from today's lecture to real financial data

# Your Portfolio Health Report

## Portfolio Health Report

Central Tendency	Mean = 8.2%, Median = 7.5%
Volatility	$\sigma = 15.3%$ annualized
Skewness	-0.4 (slight left tail)
Fat Tails	Kurtosis = 4.1 (leptokurtic)
Sharpe Ratio	0.54 (moderate)

**Diagnosis: Healthy but watch the fat tails**

---

Every dataset deserves a checkup – now you know how to give one

# Key Takeaways

## Five things to remember from this lesson:

1. **Central tendency** (mean, median) tells you where data clusters
2. **Dispersion** (std dev, IQR) tells you how spread out it is
3. **Skewness** reveals asymmetry – negative skew means crash risk
4. **Kurtosis** reveals fat tails – extreme events are more common than normal models suggest
5. **Risk metrics** (Sharpe, VaR, drawdown) translate statistics into investment decisions

---

Statistics is the language of data – fluency here unlocks everything that follows

# Next: L14 – Distributions

**Today** we described data with summary statistics.

**Next time** we model data with probability distributions:

- Normal, t, uniform, exponential distributions
- Fitting distributions to real data
- Central Limit Theorem – why normality appears everywhere
- QQ-plots: visual test for distribution fit

*Descriptive statistics summarize what happened.  
Distributions model what **could** happen.*

---

**Prepare: review normal distribution properties and the concept of probability density**