# Lesson 14: Probability Distributions

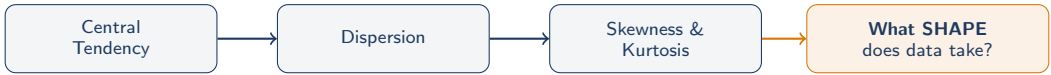
## Data Science with Python – BSc Course

Data Science Program

BSc Course

45 Minutes

## Previously on Data Science...




**Summary statistics describe data – distributions model it.**

- L13 gave us numbers: mean, standard deviation, skewness
- Today: the mathematical blueprints those numbers come from

---

Think of distributions as blueprints for randomness – they tell you what outcomes to expect

## Learning Objectives

**After this lesson, you will be able to:**

1. **Explain** the normal distribution and its key properties
2. **Compare** discrete and continuous distributions
3. **Apply** PDF and CDF to compute probabilities
4. **Analyze** fat tails and why they break financial models
5. **Evaluate** distribution fit using QQ-plots and formal tests

---

Bloom's levels: remember through evaluate – each objective builds on the previous

## Why Does the Shape of Your Data Matter?

**Because a normal distribution says crashes are impossible –**
**and they happen every decade.**


Each of these was a "once in 10,000 years" event
under the normal distribution

- October 19, 1987: S&P 500 fell 20.5% in one day
- Normal model: probability $\approx 10^{-72}$ – essentially zero
- **Choosing the wrong distribution = catastrophic risk blindness**

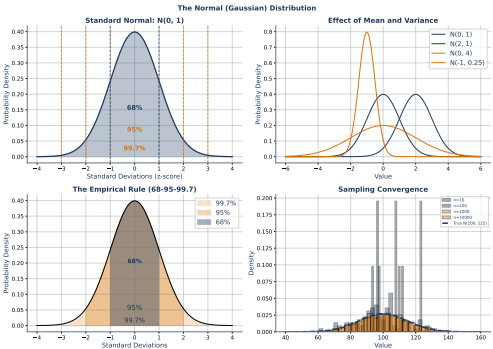
---

The normal distribution is a useful starting point – but never the final answer for financial risk

# The Normal Distribution

## The most famous distribution – and the default assumption

- Symmetric, bell-shaped, defined by mean $\mu$ and std $\sigma$
- Notation: $X \sim N(\mu, \sigma^2)$

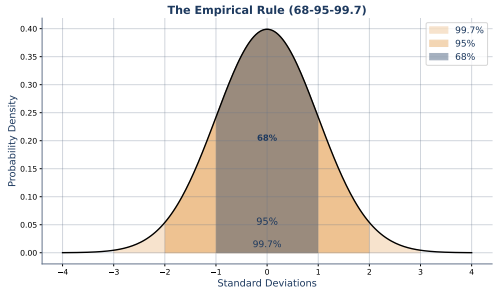



PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$

## The 68–95–99.7 Rule

**How much data falls within 1, 2, 3 standard deviations?**

- 68% within $\pm 1\sigma$, 95% within $\pm 2\sigma$, 99.7% within $\pm 3\sigma$
- Beyond $3\sigma$: only 0.3% – *if* truly normal

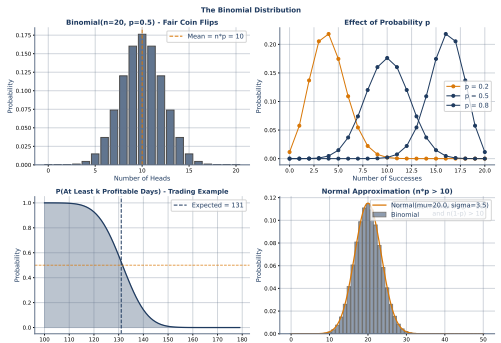



A daily return beyond $3\sigma$ should happen once every 3 years – reality: much more often

## Binomial Distribution: Counting Successes

**Discrete: how many successes in $n$ independent trials?**

- $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Example: how many of 20 trading days close positive?

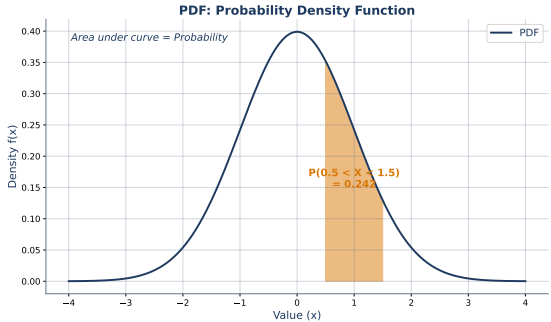



$n$ = trials, $p$ = success probability. $E[X] = np$, $Var(X) = np(1-p)$

## PDF: Probability Density Function

**For continuous variables, probability = area under the curve**

- $f(x) \geq 0$ everywhere; total area $= 1$
- $P(a \leq X \leq b) = \int_a^b f(x)\,dx$

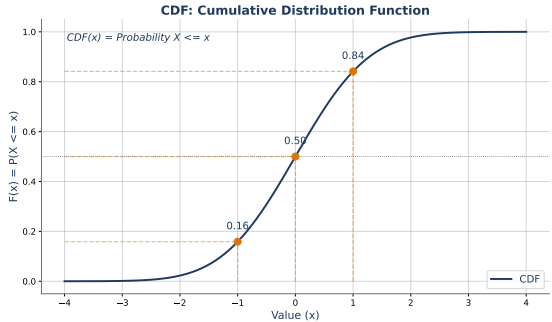



Important: $P(X = x) = 0$ for continuous variables – only intervals have probability

## CDF: Cumulative Distribution Function

$F(x) = P(X \leq x)$ – **probability of being at or below** $x$

- Monotonically increasing from 0 to 1
- $P(a < X \leq b) = F(b) - F(a)$

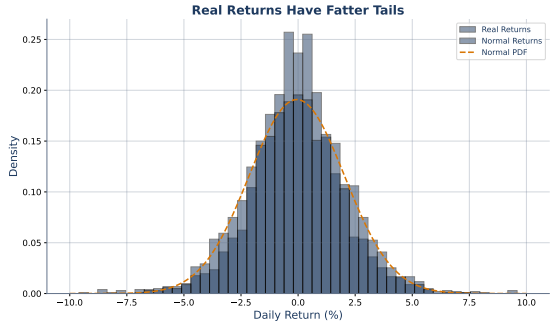



CDF = integral of PDF. In Python: `norm.cdf(x)` gives the cumulative probability

## What Real Stock Returns Look Like

### Empirical distribution of daily S&P 500 returns

- Roughly bell-shaped – but not perfectly normal
- Notice the peak is sharper and tails are heavier

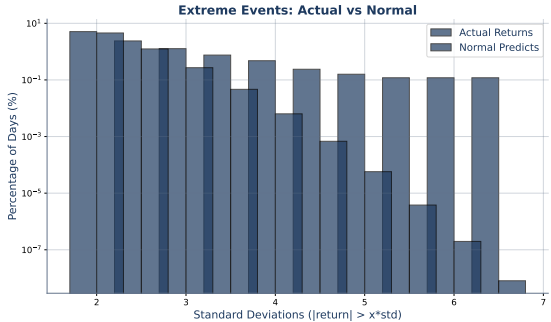



Leptokurtic: more peaked center + heavier tails than a true normal distribution

## Tails That Shouldn't Exist

### Extreme returns happen far more often than the normal predicts

- Normal says $> 4\sigma$ events are once-in-a-century rare
- Markets produce them every few years




This gap between theory and reality destroyed Long-Term Capital Management in 1998

## Checkpoint: Think About This

### Quick Check

If daily stock returns are normally distributed with $\mu = 0.04\%$ and $\sigma = 1.2\%$,

what is the probability of a $-20\%$ day?

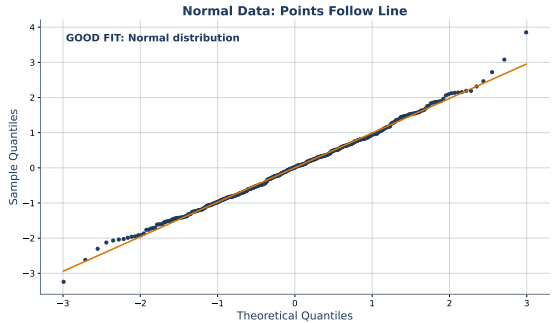
**Hint:** $-20\%$ is about 16.7 standard deviations from the mean.
Under the normal distribution: $P \approx 10^{-62}$
Yet October 19, 1987 *actually happened.*

This is why we need distributions with heavier tails than the normal

# QQ Plot: Does My Data Follow a Distribution?

## Quantile-Quantile plot compares data vs. theoretical quantiles

- Points on the line $\Rightarrow$ data matches the distribution
- Deviations reveal fat tails, skewness, or outliers

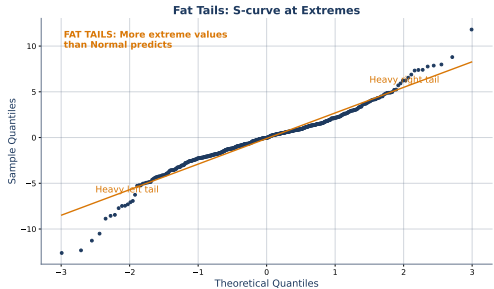



Python: `scipy.stats.probplot(data, dist="norm", plot=plt)` – fast visual diagnostic

# QQ Plot: Fat Tails Revealed

**When returns have fatter tails than normal:**

- Points curve *up* at the right, *down* at the left
- S-shaped deviation = classic fat-tail signature

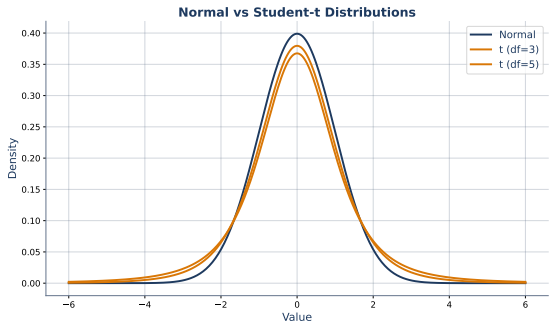



If your QQ plot shows an S-curve, the normal distribution is underestimating your tail risk

## Normal vs. Student-t Distribution

### The t-distribution: same bell shape, heavier tails

- Parameter $\nu$ (degrees of freedom) controls tail weight
- Small $\nu$ = very fat tails; as $\nu \to \infty$, t $\to$ normal
- Better fit for financial returns than the normal

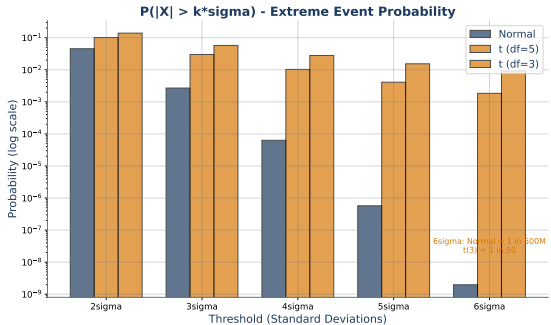



Python: `from scipy.stats import t` – use `t.pdf(x, df=5)` for fat-tailed modeling

## Extreme Event Probabilities: Normal vs. Reality

**How much do tails matter? Orders of magnitude.**

- A $3\sigma$ event: normal says 0.3%, fat-tailed says 2–5%
- A $5\sigma$ event: normal says 1 in 3.5 million, reality: far more

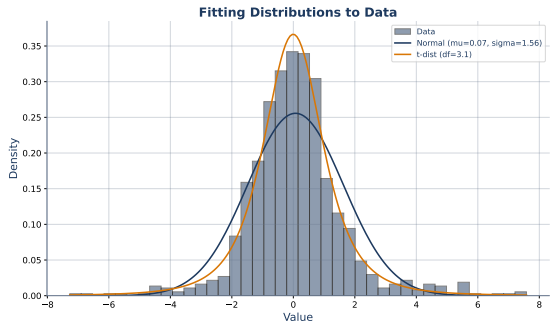



Risk managers who rely on the normal distribution systematically underestimate extreme losses

## Fitting Distributions to Data

### Overlay candidate distributions on your histogram

- Maximum Likelihood Estimation (MLE) fits parameters
- Compare: normal, t, log-normal – which fits best?

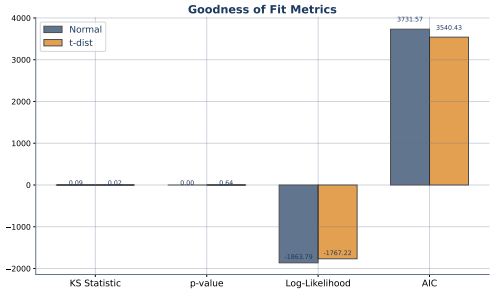



Python: `scipy.stats.norm.fit(data)` returns $(\hat{\mu}, \hat{\sigma})$ via MLE

## Goodness-of-Fit: Testing Formally

**Visual inspection is not enough – use statistical tests**

- **KS test**: max distance between empirical and theoretical CDF
- **Shapiro-Wilk**: specific to normality ($n < 5000$)
- $p < 0.05$: reject the hypothesized distribution

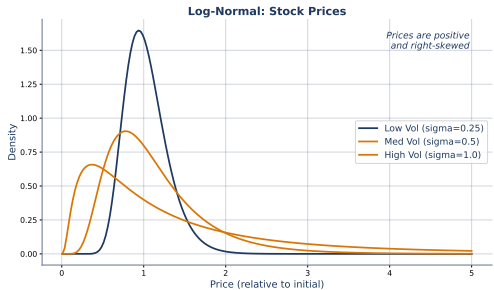



Python: `kstest(data, 'norm', args=(mu, sig))` or `shapiro(data)`

## Log-Normal: Why Prices Use It

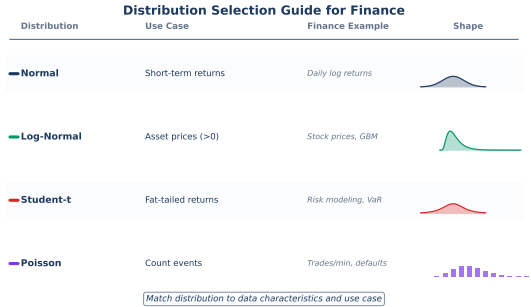
**If log-returns are normal, then prices are log-normal**

- Prices can never go negative (log-normal $> 0$ always)
- Foundation of Black-Scholes option pricing




Key insight: model returns as normal, then exponentiate to get prices

# Distributions in Finance: Which One When?




- **Normal**: short-term returns / **t-distribution**: fat-tailed risk
- **Log-normal**: stock prices / **Poisson**: event counts

Match the distribution to your data type – there is no single "right" distribution

## Hands-On: Is Your Data Normal?

**Exercise (5 min):** Load real stock returns and test normality.

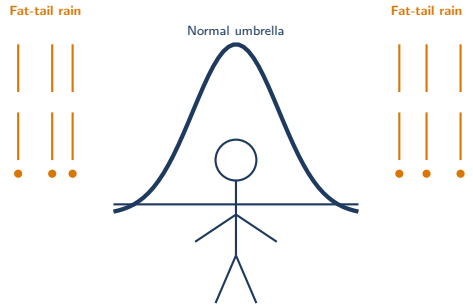
1. Generate 500 daily returns: `np.random.normal(0.0004, 0.012, 500)`
2. Plot histogram with `plt.hist(returns, bins=40, density=True)`
3. Create QQ plot: `stats.probplot(returns, plot=plt)`
4. Run Shapiro-Wilk: `stat, p = stats.shapiro(returns)`
5. Interpret: is $p < 0.05$? What does the QQ plot show?

**Bonus:** Replace `np.random.normal` with `stats.t.rvs(df=5, size=500)`.
How do the results change?

---

This exercise combines histogram, QQ plot, and formal testing – the three tools for distribution analysis

# The Moral of the Story




The normal distribution protects
the center but misses the extremes

A normal-distribution umbrella leaves you exposed to exactly the events that matter most

## Key Takeaways

**What you learned today:**

1. **Distributions are blueprints** for random processes – they predict what to expect
2. **Normal distribution** is defined by $\mu$ and $\sigma$; 68-95-99.7 rule applies
3. **PDF gives density**, CDF gives cumulative probability – both are essential tools
4. **Real financial returns have fat tails** – the normal underestimates extremes
5. **QQ plots and formal tests** (KS, Shapiro-Wilk) assess distribution fit
6. **Choose the right model**: normal for quick estimates, t for risk, log-normal for prices

---

Key message: never assume normality in finance – always check, always test

**Now that you know distributions, you can test claims about data.**

- Is the average return *really* positive, or just noise?
- Did a strategy *actually* outperform the market?
- Are two portfolios *significantly* different?

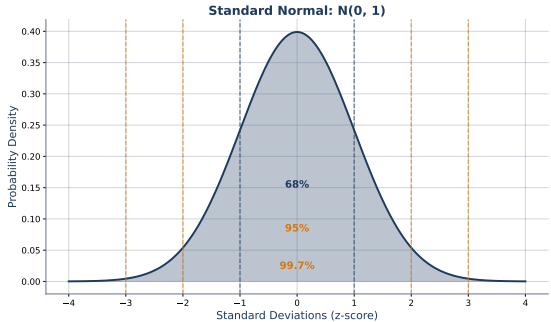
---

L15 builds directly on today's material – distributions are the engine behind every hypothesis test

## Self-Study: Standard Normal Distribution

**The standard normal:** $Z \sim N(0, 1)$

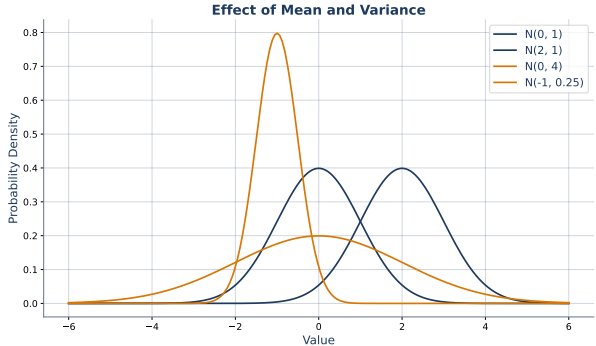
- Any normal variable can be standardized: $Z = (X - \mu)/\sigma$
- Z-tables give probabilities for standard normal
- Python: `norm.cdf(z)` for any z-score




Standardization lets you compare variables with different units and scales

## Self-Study: How Mean and Variance Shape the Curve

### $\mu$ shifts the center; $\sigma$ controls the spread

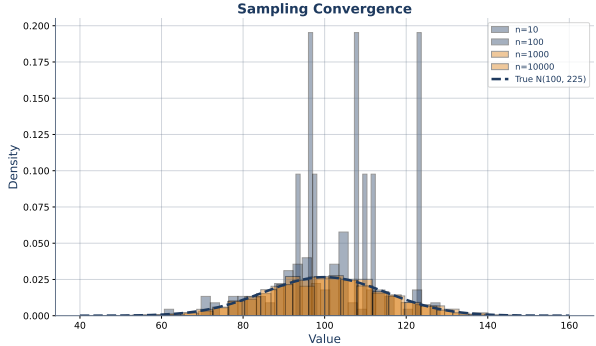



- Larger $\sigma$: wider, shorter bell curve (same total area)
- Different $\mu$: same shape, different location

Two parameters fully define a normal distribution – this simplicity is both its strength and limitation

## Self-Study: Central Limit Theorem

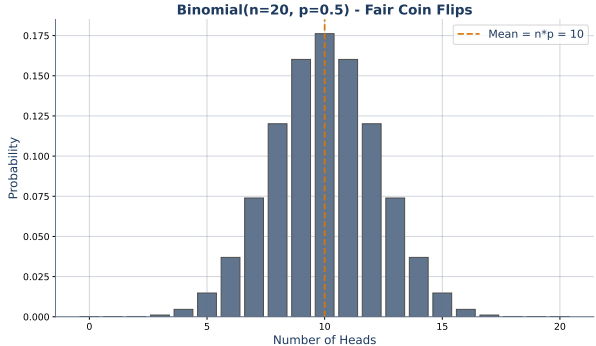
### Sample means converge to normal regardless of original shape




- $\bar{X}_n \xrightarrow{d} N(\mu, \sigma^2/n)$ as $n \to \infty$
- Practical rule: works well for $n \geq 30$
- This is why normal appears everywhere in statistics

CLT is the mathematical reason the normal distribution is so central to data science

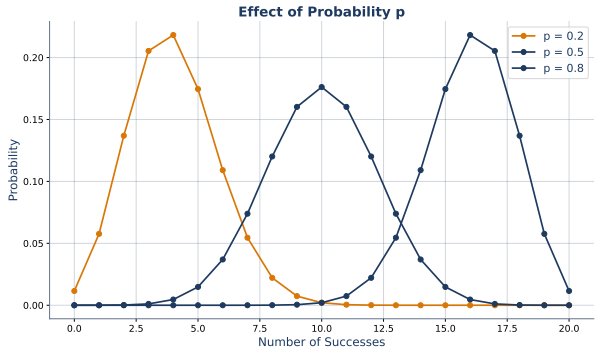
## Self-Study: Binomial – Coin Flip Example




- $n = 10$ flips, $p = 0.5$: symmetric distribution
- Most likely outcome: 5 heads (but 4 or 6 are nearly as likely)

The binomial is the simplest discrete distribution – foundation for counting problems

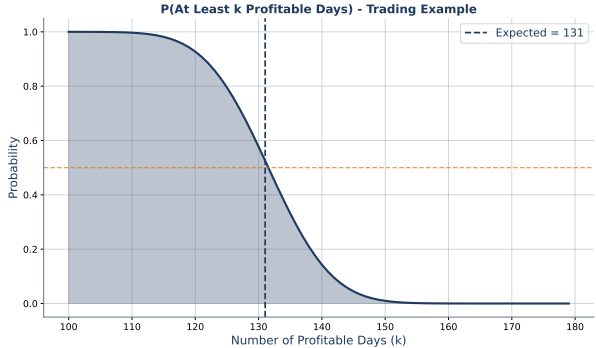
## Self-Study: Binomial – Different Success Probabilities




- $p = 0.5$: symmetric; $p < 0.5$: left-skewed; $p > 0.5$: right-skewed
- Finance: $p \approx 0.53$ for daily positive returns historically

Changing $p$ shifts and skews the binomial distribution

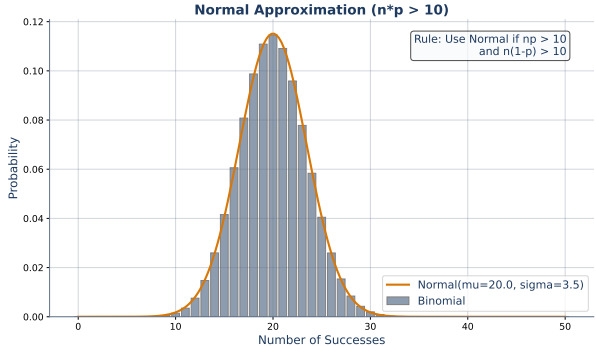
## Self-Study: Binomial – Positive Trading Days




- Model: $X \sim Bin(n = 22, p = 0.53)$ for one month of trading
- Expected positive days: $np = 22 \times 0.53 \approx 11.7$

Binomial models give quick probability estimates for up/down day counts

## Self-Study: Binomial – Normal Approximation

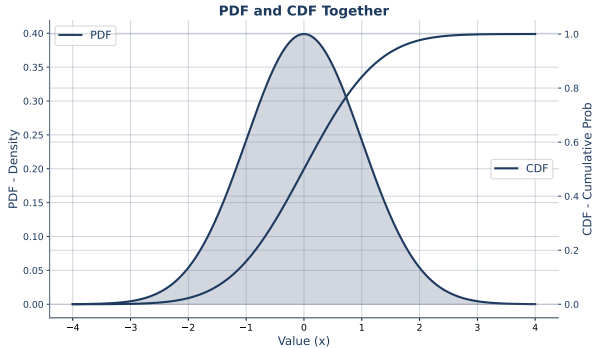



- For large $n$: $Bin(n, p) \approx N(np, np(1-p))$
- Rule of thumb: works when $np \geq 5$ and $n(1-p) \geq 5$

---

The normal approximation to the binomial is another consequence of the Central Limit Theorem

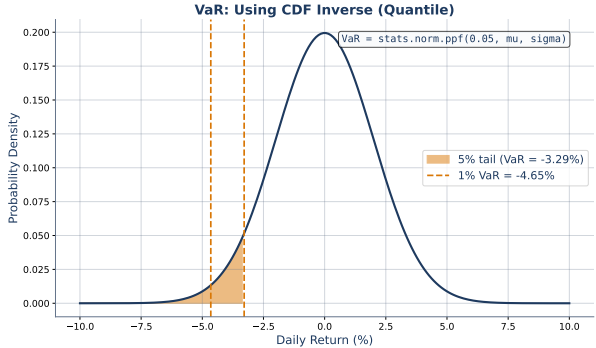
## Self-Study: PDF and CDF Together




- CDF is the integral of the PDF: $F(x) = \int_{-\infty}^{x} f(t)\,dt$
- PDF is the derivative of the CDF: $f(x) = F'(x)$
- CDF always increases from 0 to 1

Understanding the PDF-CDF relationship is essential for probability calculations

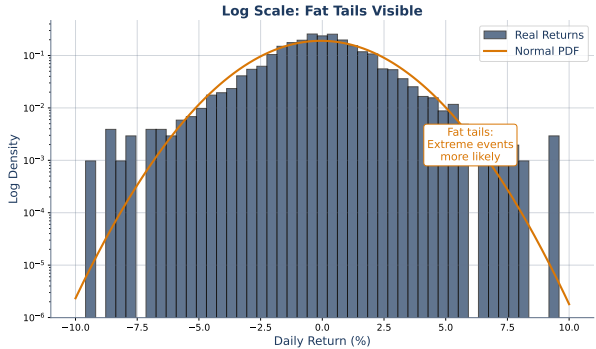
# Self-Study: Value at Risk from Quantiles




- VaR at $\alpha = 5\%$: the 5th percentile of the loss distribution
- Python: `norm.ppf(0.05, loc=mu, scale=sigma)`
- Interpretation: "95% of the time, losses won't exceed this"

VaR is one of the most widely used risk measures in banking – and it depends on your choice of distribution

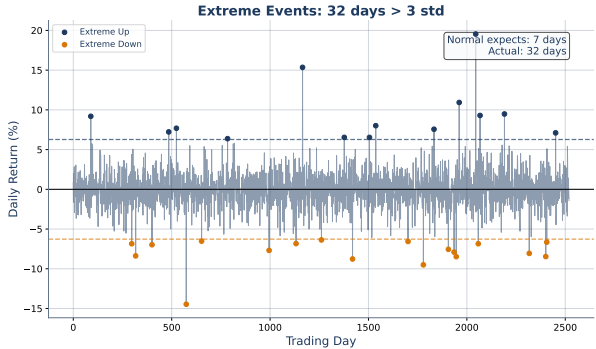
## Self-Study: Histogram on Log Scale




- Log scale reveals tail behavior that linear scale hides
- A normal distribution appears as a parabola on log scale
- Heavier tails appear as flatter slopes

Log-scale histograms are essential for diagnosing tail heaviness

## Self-Study: Returns Over Time



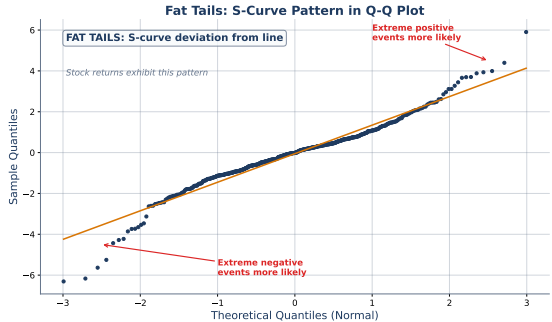

- Returns cluster: large moves follow large moves (volatility clustering)
- This violates the i.i.d. assumption of simple distribution models
- More advanced: GARCH models capture time-varying volatility

Distributions describe the marginal behavior – time series models capture the temporal structure

## Self-Study: Reading Fat Tails in QQ Plots

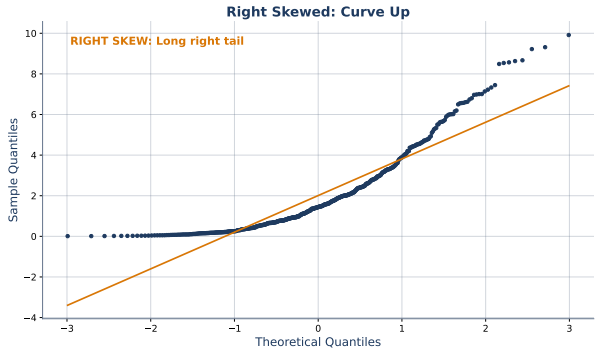
**Pattern recognition guide for QQ plot interpretation:**

- **Points on line**: data matches the reference distribution
- **S-curve (tails bend away)**: fat tails (leptokurtic)
- **Reverse S**: thin tails (platykurtic)
- **Single curve**: skewed distribution
- **Steps/jumps**: discrete data or rounding




QQ plots work for any reference distribution – not just the normal

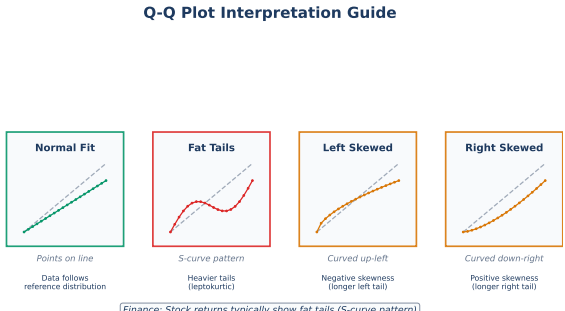
# Self-Study: QQ Plot – Right-Skewed Data




- Right-skewed data curves upward from the reference line
- Common in finance: income, wealth, company sizes
- Consider log-normal or gamma distribution instead

Skewness in QQ plots suggests you need an asymmetric distribution model

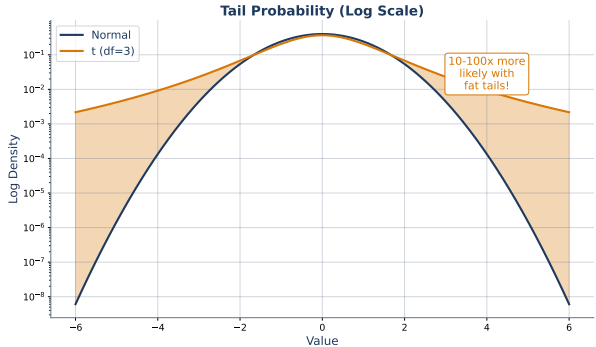
# Self-Study: QQ Plot Interpretation Summary




- Always compare against a relevant theoretical distribution
- Use `scipy.stats.probplot(data, dist="t", sparams=(5,))` for t-distribution QQ

QQ plots are the single most useful visual tool for distribution assessment

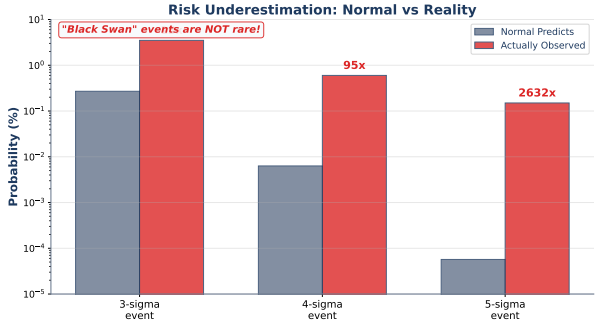
## Self-Study: Fat Tails on Log Scale




- Log-scale comparison makes tail differences dramatic
- Normal tails drop off exponentially; t-distribution tails drop as power law
- Power-law tails assign much higher probability to extreme events

The log-scale view is how quants diagnose tail risk – linear scale hides it

# Self-Study: Fat Tails – Investment Implications

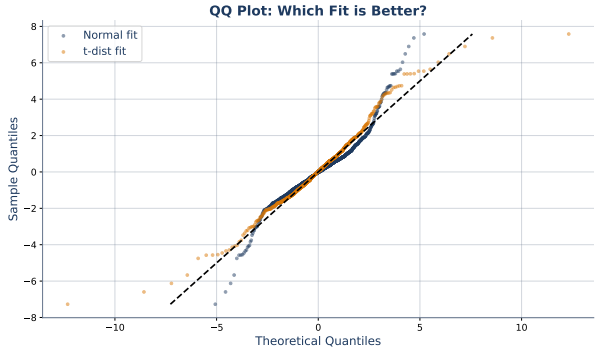



*VaR models using normal assumptions significantly underestimate tail risk*

- Normal-based VaR underestimates losses by 2–10x in the tails
- Black swan events are more likely than models predict
- Solutions: t-distribution, historical simulation, extreme value theory

Risk models must account for fat tails – regulators increasingly require it

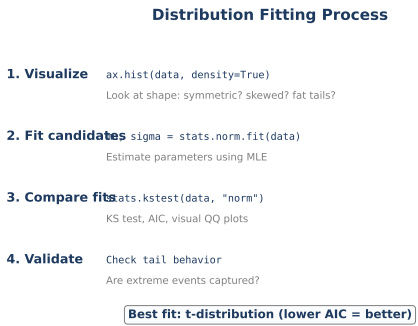
## Self-Study: QQ Plots for Multiple Distributions




- Compare QQ plots against normal, t, and log-normal
- The distribution whose QQ plot is closest to the line wins

Combining QQ plots with formal tests gives the most robust distribution selection

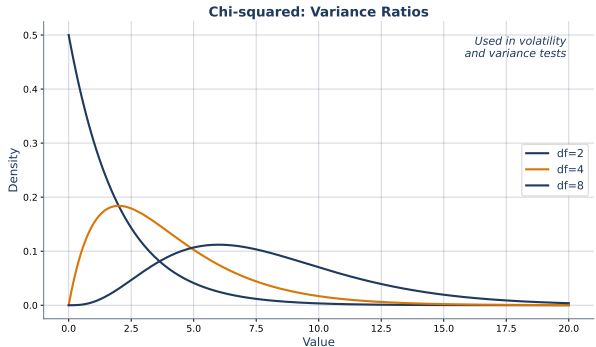
# Self-Study: Distribution Fitting Process




- Step 1: Visualize (histogram + density overlay)
- Step 2: Fit candidate distributions (MLE)
- Step 3: Test (KS, Anderson-Darling, AIC)
- Step 4: Select best model, validate on held-out data

This four-step process applies to any distribution fitting problem

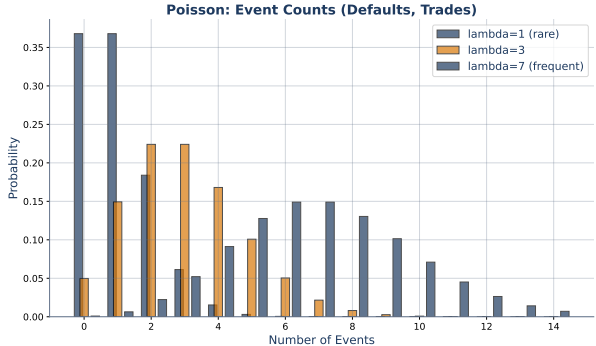
## Self-Study: Chi-Squared Distribution




- $\chi^2_k$ = sum of $k$ squared standard normals
- Used in variance testing and goodness-of-fit tests
- Always positive, right-skewed (becomes more symmetric as $k$ grows)

You will encounter chi-squared in hypothesis testing (L15) and categorical data analysis

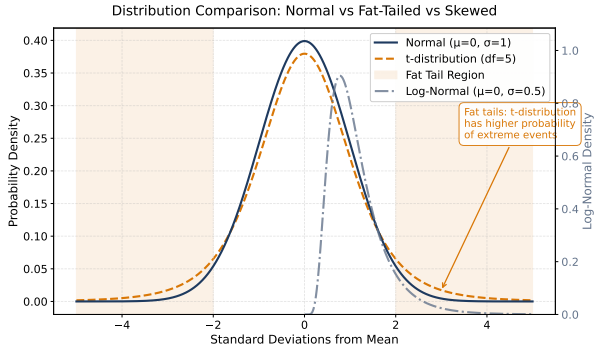
## Self-Study: Poisson Distribution




- Counts events in fixed intervals: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- Finance: number of trades per minute, defaults per year
- Single parameter $\lambda$ = mean = variance

Poisson is the go-to distribution for rare event counting in finance

# Self-Study: Distribution Comparison Overview




- Normal: symmetric, light tails – baseline for comparison
- Student-t: symmetric, heavy tails – better for returns
- Log-normal: right-skewed, positive only – prices, wealth
- Poisson: discrete, counts only – events per interval

Choosing the right distribution is about matching mathematical properties to your data's characteristics