# Lesson 14: Probability Distributions

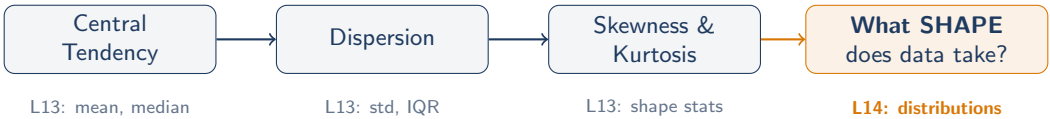
## Data Science with Python – BSc Course

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:**

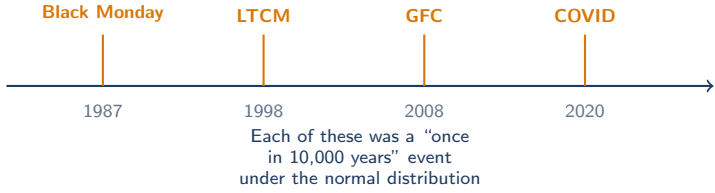
1. **Explain** the normal distribution and its key properties
2. **Compare** discrete and continuous distributions
3. **Apply** PDF and CDF to compute probabilities
4. **Analyze** fat tails and why they break financial models
5. **Evaluate** distribution fit using QQ-plots and formal tests

---

**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.**




- 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**

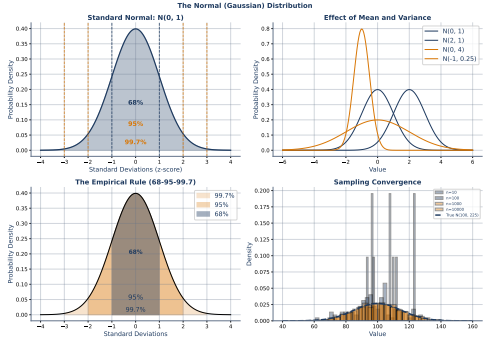
---

**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: $X \sim N(\mu, \sigma^2)$

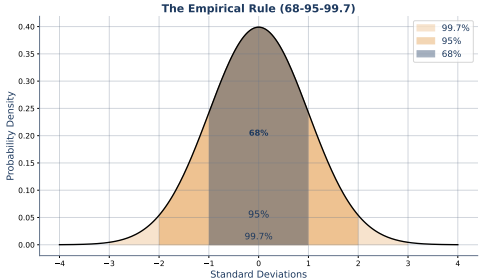



**PDF:** $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$

# 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

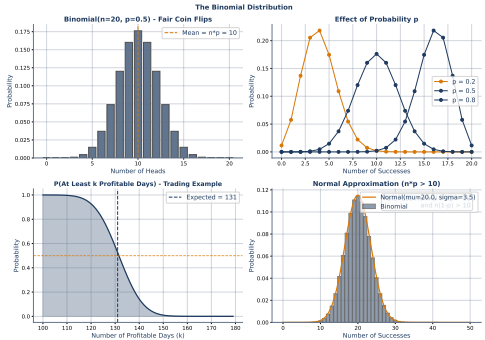



**A daily return beyond $3\sigma$ should happen once every 3 years – reality: much more often**

# 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?

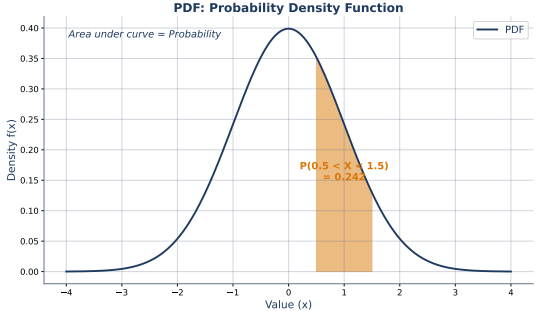



$n$ = **trials,** $p$ = **success probability.** $E[X] = np$, $Var(X) = np(1-p)$

# PDF: Probability Density Function

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

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

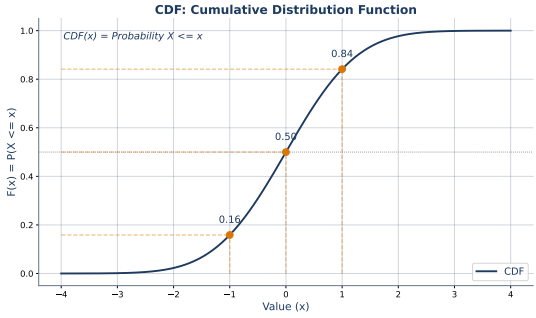



**Important:** $P(X = x) = 0$ **for continuous variables – only intervals have probability**

# CDF: Cumulative Distribution Function

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

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

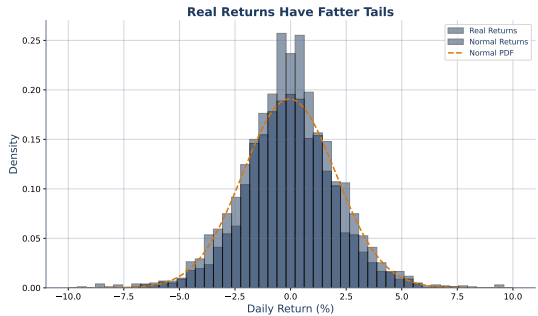



**CDF = integral of PDF. In Python:** `norm.cdf(x)` **gives the cumulative probability**

# 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

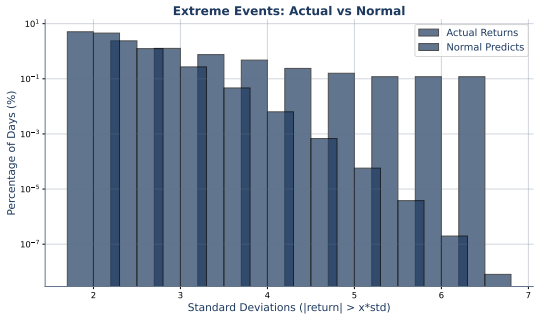



**Leptokurtic: more peaked center + heavier tails than a true normal distribution**

# 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: $P \approx 10^{-62}$

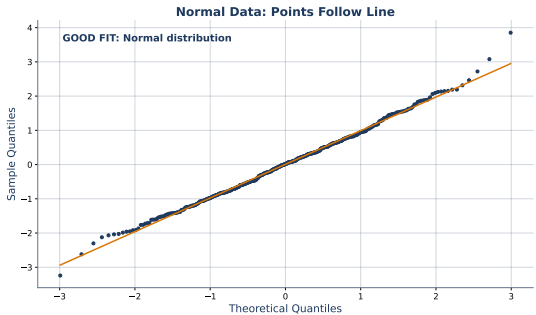
Yet October 19, 1987 *actually happened*.

**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

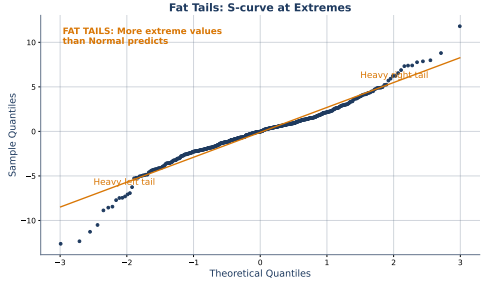



**Python:** `scipy.stats.probplot(data, dist="norm", plot=plt)` – **fast visual diagnostic**

# 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

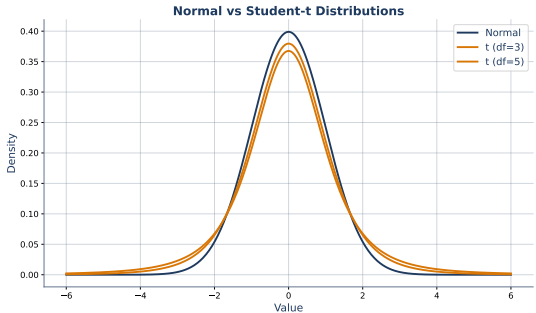



**If your QQ plot shows an S-curve, the normal distribution is underestimating your tail risk**

# 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

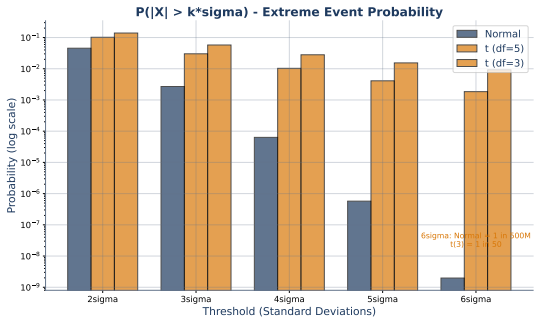



**Python:** `from scipy.stats import t` – **use** `t.pdf(x, df=5)` **for fat-tailed modeling**

# 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

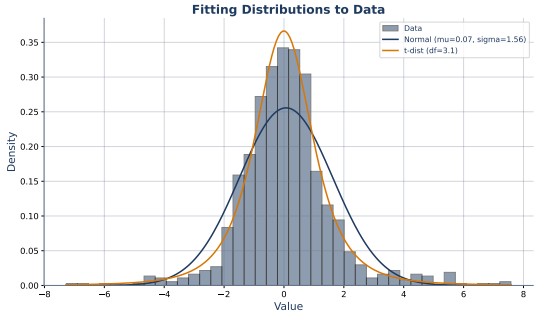



**Risk managers who rely on the normal distribution systematically underestimate extreme losses**

# Fitting Distributions to Data

## Overlay candidate distributions on your histogram

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

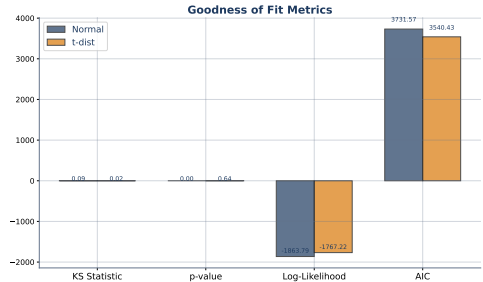



**Python:** `scipy.stats.norm.fit(data)` **returns** $(\hat{\mu}, \hat{\sigma})$ **via MLE**

# 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

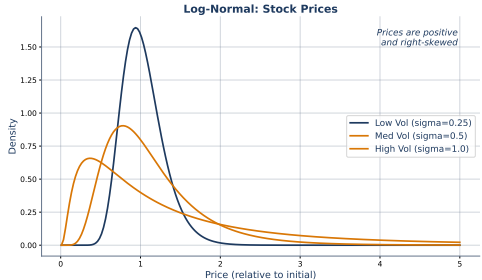



**Python:** `kstest(data, 'norm', args=(mu, sig))` **or** `shapiro(data)`

# Log-Normal: Why Prices Use It

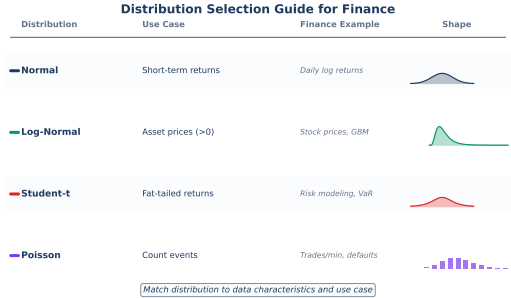
**If log-returns are normal, then prices are log-normal**

- 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.

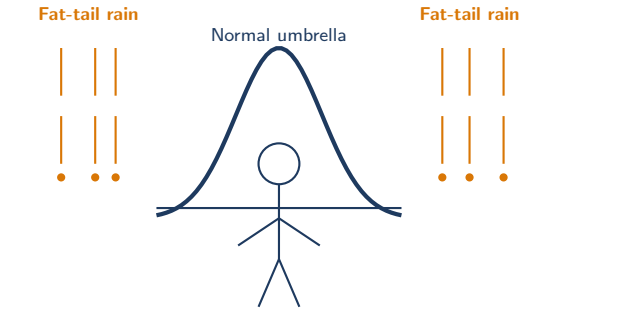
1. Generate 500 daily returns: `np.random.normal(0.0004, 0.012, 500)`
2. Plot histogram with `plt.hist(returns, bins=40, density=True)`
3. Create QQ plot: `stats.probplot(returns, plot=plt)`
4. Run Shapiro-Wilk: `stat, p = stats.shapiro(returns)`
5. Interpret: is $p < 0.05$? What does the QQ plot show?

**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




**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**