Lesson 15: Hypothesis Testing

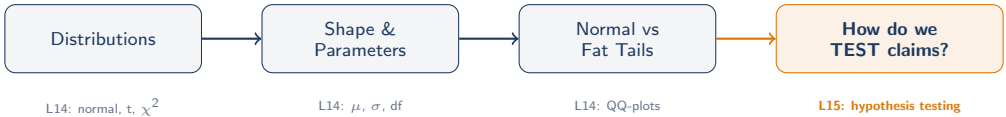
Data Science with Python – BSc Course

Data Science Program

BSc Course

45 Minutes

Previously on Data Science...



Distributions model randomness – hypothesis testing judges it.

- L14 gave us the shapes data can take
- Today: a rigorous framework for deciding if patterns are real

---

Think of hypothesis testing as putting your data on trial

Learning Objectives

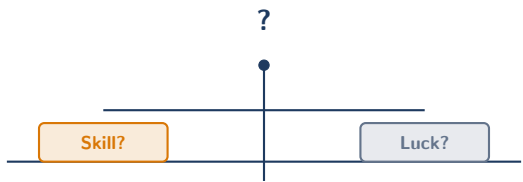
After this lesson, you will be able to:

1. **Explain** null and alternative hypotheses (H_0 vs H_1)
2. **Apply** one-sample and two-sample t-tests in Python
3. **Interpret** p-values correctly (and avoid common traps)
4. **Analyze** Type I and Type II errors and their trade-offs
5. **Design** A/B tests with appropriate sample sizes

Bloom's levels: explain, apply, interpret, analyze, design

A hedge fund claims 12% annual returns.

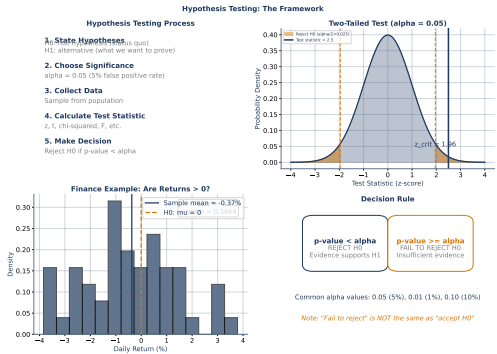
Is this skill or luck?



How much evidence do we need to decide?

Hypothesis testing provides a principled framework for this exact question

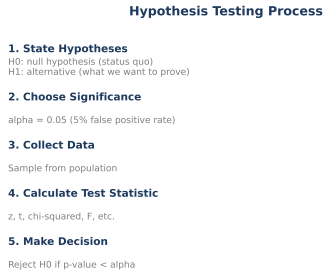
The Logic of Hypothesis Testing



- Assume H_0: "nothing unusual is happening"
- Collect data and measure how surprising it is under H_0
- Very unlikely data under $H_0 \rightarrow$ reject H_0

Like a courtroom: innocent until proven guilty beyond reasonable doubt

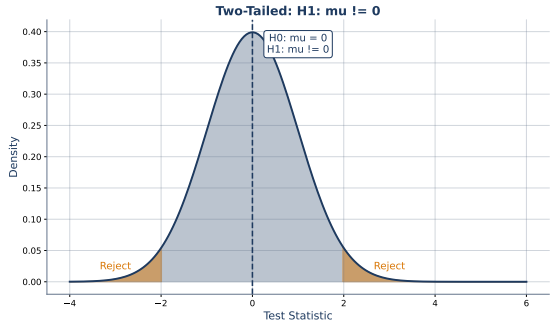
The 5-Step Testing Process



1. State hypotheses: H_0 (null) vs H_1 (alternative)
2. Choose significance level: $\alpha = 0.05$ typical
3. Collect data from population
4. Calculate test statistic (t, z, F, χ^2)
5. Decide: reject H_0 if p-value $< \alpha$

This 5-step process applies to every hypothesis test you will encounter

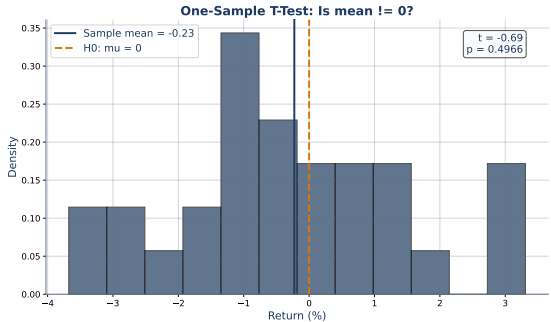
Rejection Regions: Two-Tailed Test



- H_0: $\mu = \mu_0$ vs H_1: $\mu \neq \mu_0$
- Rejection in **both** tails – evidence for difference in either direction
- Each tail gets $\alpha/2$ of the significance level

Two-tailed is the default – use it unless you have a strong directional hypothesis

One-Sample t-Test

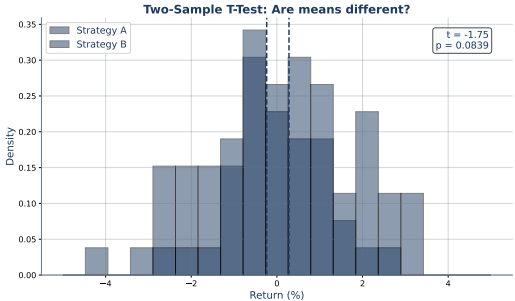


Formula: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad df = n - 1$

When to use: Test whether a sample mean differs from a hypothesized value.
Example: Is the mean daily return of a stock significantly different from zero?

The t-statistic measures how many standard errors $\bar{x}$ is from μ_0

Two-Sample t-Test (Welch's)

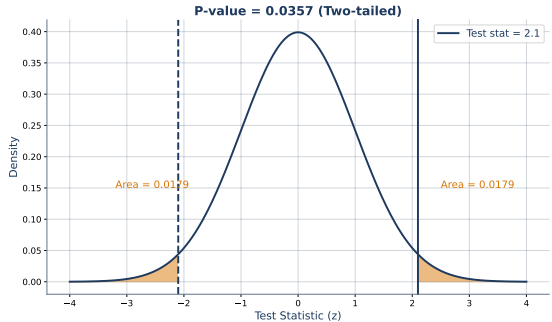


Formula: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$

When to use: Compare means of two independent groups.
Example: Do tech stocks have higher returns than utilities?

Welch's t-test does NOT assume equal variances – use it by default in Python

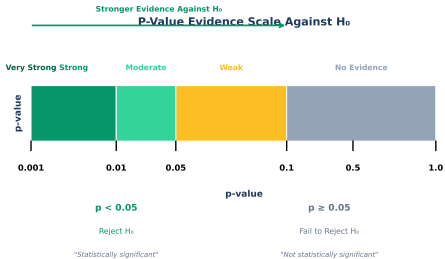
What Is a p-Value?



- **Definition:** Probability of observing data this extreme *if* H_0 *is true*
- Small p-value $\rightarrow$ strong evidence against H_0
- Decision rule: reject H_0 if p-value $< \alpha$

The p-value is the shaded tail area beyond the observed test statistic

p-Value: What It Means (and Doesn't Mean)



p-value IS: probability of data this extreme assuming H_0 is true

p-value is NOT:

- The probability that H_0 is true
- The probability results are "due to chance"
- A measure of effect size (small p $\neq$ large effect)

This distinction trips up even experienced researchers – memorize it

Quick Check

You test a trading strategy and get $p = 0.03$.
Do you reject or fail to reject H_0 at $\alpha = 0.05$?

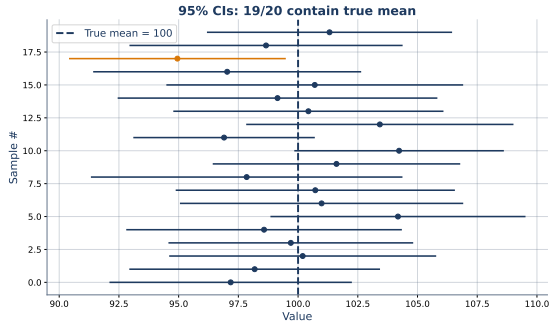
Answer

$p = 0.03 < \alpha = 0.05 \Rightarrow$ **Reject** H_0

- Evidence is strong enough to reject the null hypothesis
- But remember: statistical significance $\neq$ practical significance

Always pair rejection with effect size – is the signal large enough to trade on?

Confidence Intervals



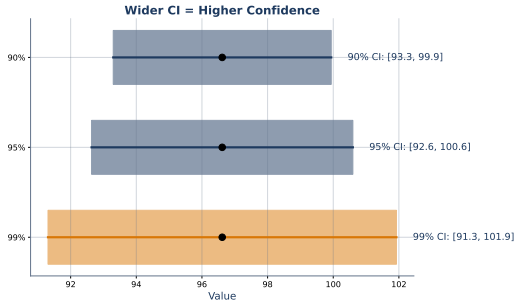
Formula: $\text{CI} = \bar{x} \pm t_{\alpha/2} \cdot \dfrac{s}{\sqrt{n}}$

- A 95% CI means: if we repeated sampling 100 times, ≈95 intervals would contain the true μ
- If μ_0 falls outside the CI $\rightarrow$ reject H_0 at that confidence level

---

Confidence intervals and hypothesis tests are two sides of the same coin

Confidence Levels: 90%, 95%, 99%

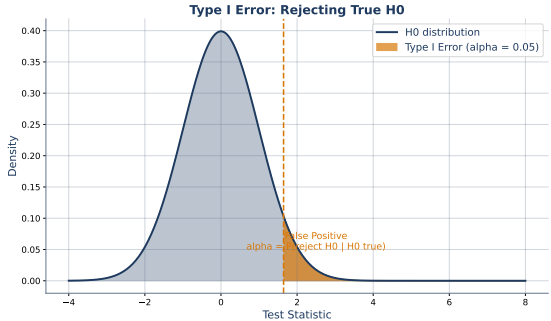


- Higher confidence → wider interval (more uncertainty)
- 95% is the standard in most research
- Trade-off: precision vs certainty

Python: `stats.t.interval(0.95, df=n-1, loc=mean, scale=se)`

Wider intervals are more likely to contain the truth – but less informative

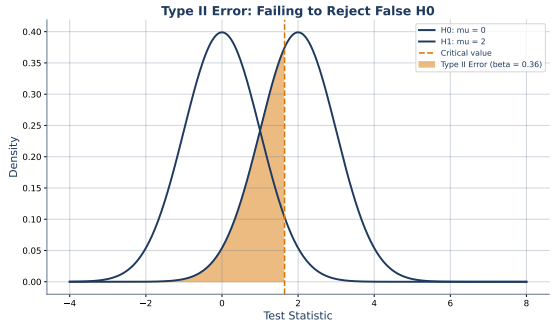
Type I Error: False Positive



- **Definition:** Rejecting H_0 when it is actually true
- **Probability:** α (the significance level you chose)
- **Finance:** Deploying a strategy that has no real alpha

Type I = false alarm. You convicted an innocent defendant

Type II Error: False Negative



- **Definition:** Failing to reject H_0 when it is actually false
- **Probability:** β **Power** $= 1 - \beta$
- **Finance:** Missing a profitable strategy due to insufficient data

---

Type II = missed detection. You let a guilty defendant walk free

Error Decision Matrix

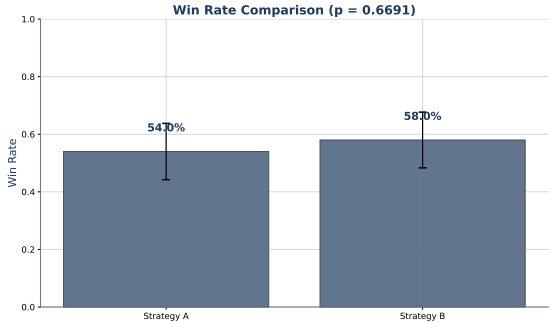
Decision Matrix

	H0 True	H0 False
Reject H0	Type I Error (False Positive) alpha	Correct! (True Positive) 1 - beta (Power)
Fail to Reject	Correct! (True Negative) 1 - alpha	Type II Error (False Negative) beta

- Lowering α (fewer false positives) $\rightarrow$ increases β (more false negatives)
- You cannot minimize both simultaneously
- Convention: $\alpha = 0.05$, target power ≥ 0.80

Every test is a trade-off between these two types of error

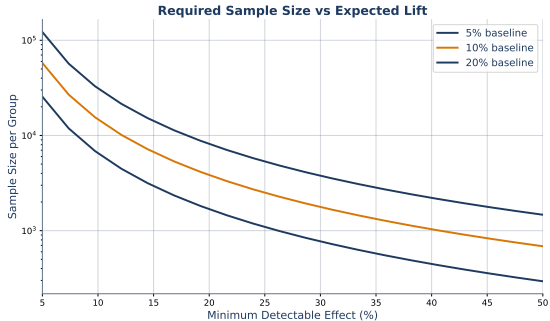
A/B Testing: Comparing Strategies



- Strategy A (control) vs Strategy B (treatment)
- Use a two-sample t-test to compare mean returns
- Randomly assign to groups to avoid confounding

A/B testing applies hypothesis testing to real-world decision making

How Much Data Do You Need?

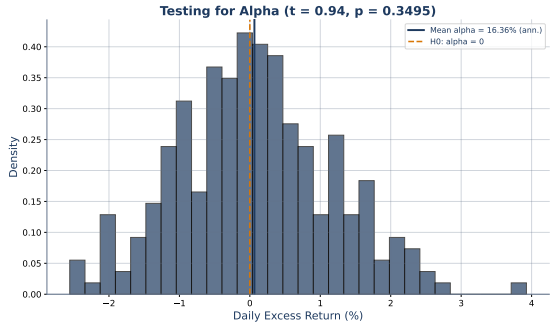


- More data → higher power → detect smaller effects
- Underpowered studies waste resources and miss real signals
- **Rule of thumb:** run a power analysis *before* collecting data

Python: `from statsmodels.stats.power import TTestIndPower`

Planning sample size upfront prevents both wasted effort and missed discoveries

Finance Application: Testing Fund Alpha



- H_0: Fund alpha $= 0$ (no skill, just market exposure)
- H_1: Fund alpha $\neq 0$ (genuine outperformance)
- Most hedge funds fail to reject H_0 after fees

Jensen's alpha test: the ultimate skill-vs-luck filter in finance

Hands-On Exercise

Test whether Apple's mean daily return differs from zero:

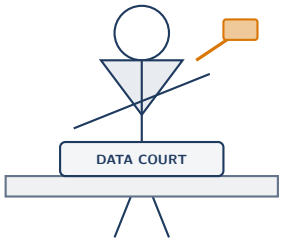
1. Load AAPL daily returns (use `yfinance` or sample data)
2. State hypotheses: H_0: $\mu = 0$ $\quad$ H_1: $\mu \neq 0$
3. Run one-sample t-test:
   ```
   from scipy import stats
   t_stat, p_val = stats.ttest_1samp(returns, 0)
   ```
4. Report: t-statistic, p-value, and 95% confidence interval
5. Interpret: reject or fail to reject at $\alpha = 0.05$?

Bonus: Compare AAPL vs MSFT returns with a two-sample t-test.
Is the difference statistically significant?

---

5 minutes – apply the full 5-step process from slide 6

The Verdict



The Verdict

- H_0 is **innocent until proven guilty**
- The p-value measures **strength of evidence**
- "Not guilty" $\neq$ "innocent"

Failing to reject H_0 does not prove H_0 is true.
It means the evidence was not strong enough.

The courtroom analogy: absence of evidence is not evidence of absence

Key Takeaways

Five Things to Remember:

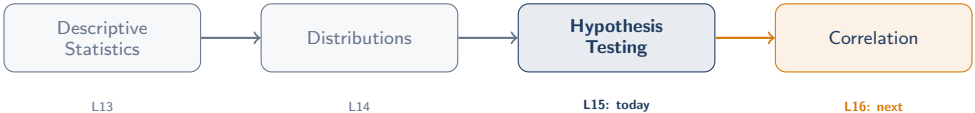
1. **Hypotheses come first:** formulate H_0 and H_1 before touching data
2. **p-value** $=$ probability of data this extreme *if* H_0 is true (not the reverse!)
3. **t-tests** compare means: one-sample (vs value), two-sample (vs group)
4. **Two error types:** false positive (α) vs false negative (β) – you cannot minimize both
5. **Effect size matters:** a tiny p-value with a tiny effect is not actionable

Python toolkit:

`scipy.stats.ttest_1samp()` `scipy.stats.ttest_ind()`
`statsmodels.stats.power.TTestIndPower()`

Statistical significance is the starting point – practical significance is the goal

Next: Correlation (L16)

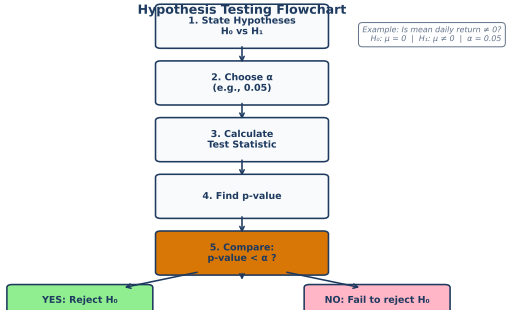


Coming up in L16:

- Pearson and Spearman correlation coefficients
- Correlation vs causation (the most misunderstood concept in data science)
- Spurious correlations and confounding variables
- Portfolio diversification: why correlation drives risk

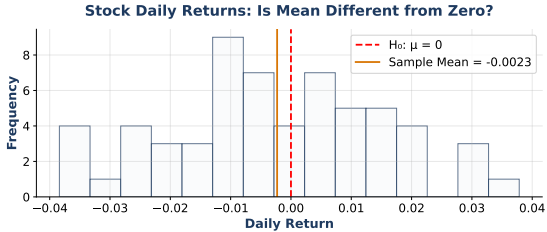
We tested single claims – next we measure how variables move together

Self-Study: Hypothesis Testing Flowchart



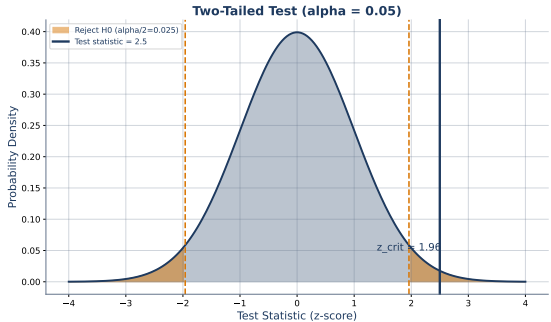
Self-study: visual roadmap of the complete testing process

Self-Study: Concrete Example with Stock Returns



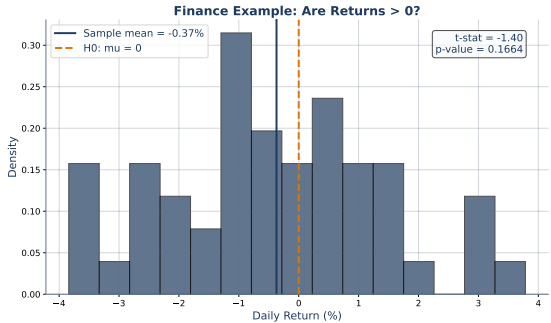
Self-study: follow the 5-step process with real data

Self-Study: Hypothesis Testing Visual



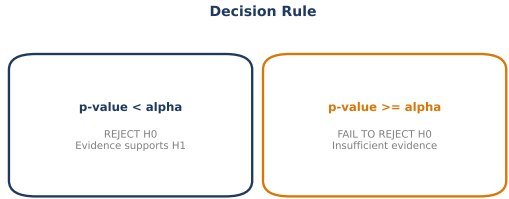
Self-study: rejection region determined by significance level alpha

Self-Study: Finance Hypothesis Example



Self-study: H0: mean return = 0 vs H1: mean return != 0

Self-Study: Decision Rule Details

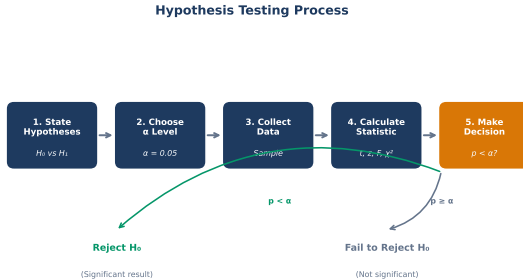


Common alpha values: 0.05 (5%), 0.01 (1%), 0.10 (10%)

Note: "Fail to reject" is NOT the same as "accept H0"

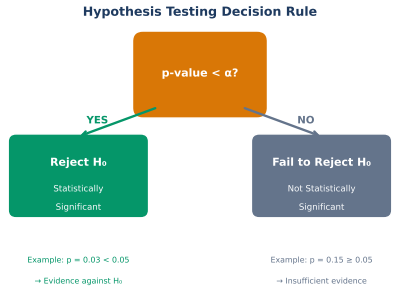
Self-study: reject H0 if p-value ¡ alpha

Self-Study: Process Flowchart



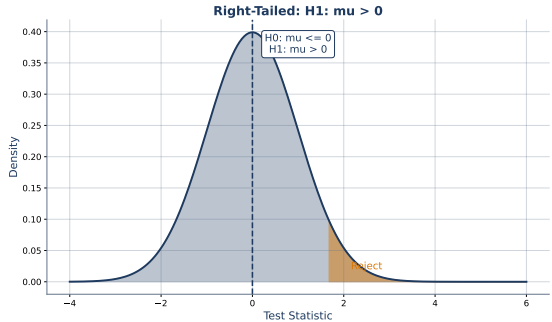
Self-study: detailed hypothesis testing process

Self-Study: Decision Rule Expanded



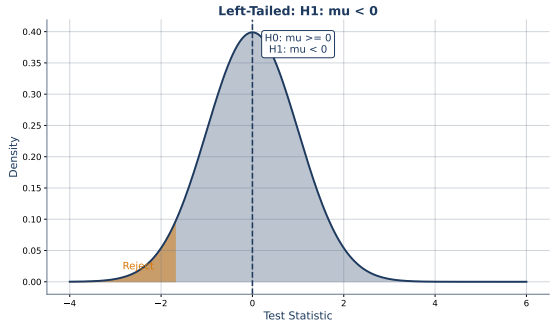
Self-study: common alpha levels and their interpretation

Self-Study: Right-Tailed Test



Self-study: H1: parameter ¿ value; reject in right tail only

Self-Study: Left-Tailed Test



Self-study: H1: parameter ¡ value; reject in left tail only

Self-Study: Hypothesis Examples

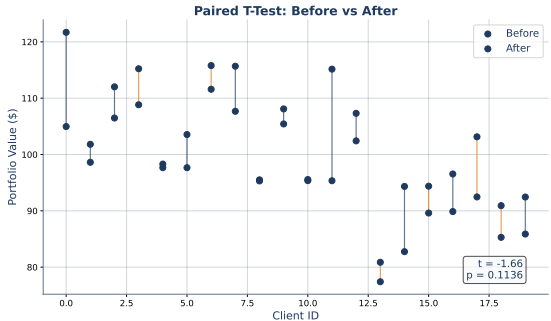
Finance Hypothesis Examples

Strategy alpha:	H0: alpha = 0 (no excess return) H1: alpha > 0 (positive alpha)	Right-tailed
Market efficiency:	H0: returns are random H1: returns are predictable	Two-tailed
Risk reduction:	H0: volatility unchanged H1: volatility decreased	Left-tailed
Correlation change:	H0: rho1 = rho2 H1: rho1 != rho2	Two-tailed

Choose test direction based on what you want to prove!

Self-study: choose test direction based on research question

Self-Study: Paired t-Test



Self-study: before-after or matched pairs comparison

Self-Study: t-Test Type Summary

T-Test Summary

One-Sample: `stats.ttest_1samp(sample, mu0)`
Compare sample mean to known value

Two-Sample: `stats.ttest_ind(sample1, sample2)`
Compare means of two groups

Paired: `stats.ttest_rel(before, after)`
Compare paired observations

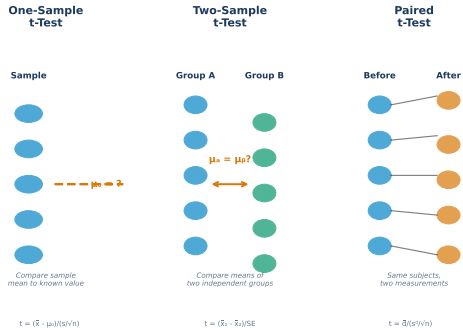
Welch: `stats.ttest_ind(..., equal_var=False)`
Unequal variances

Assumptions: Normal distribution (or large n > 30)

For non-normal: use Mann-Whitney U or Wilcoxon tests

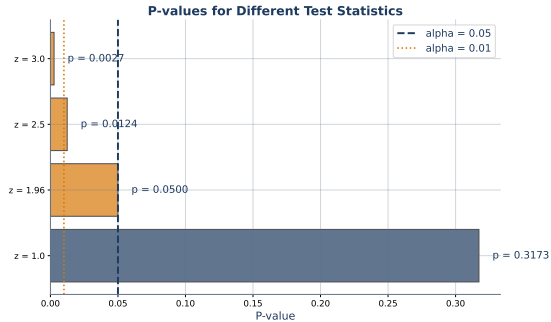
Self-study: one-sample, two-sample, and paired t-tests compared

Self-Study: t-Test Comparison



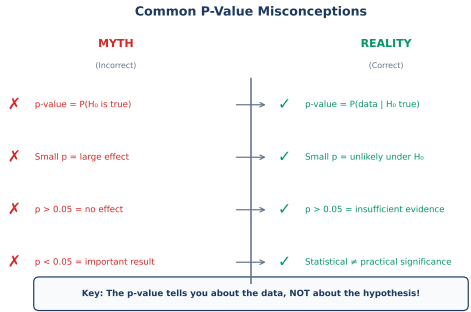
Self-study: visual comparison of all three t-test types

Self-Study: p-Values for Different Statistics



Self-study: same logic applies to z, t, F, chi-square tests

Self-Study: Common p-Value Mistakes



Self-study: avoid these common misinterpretations

Self-Study: p-Value Mistakes Expanded

P-Value: What It Is NOT

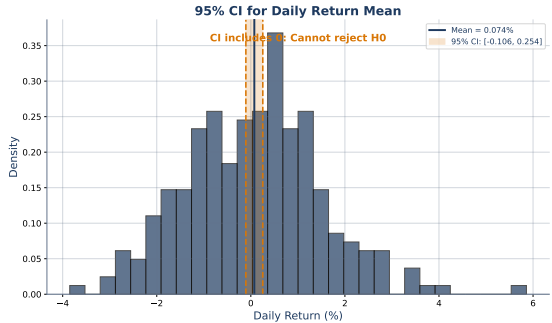
- X Probability that H0 is true
- X Probability that H1 is true
- X Probability of making an error
- X Size of the effect
- X Importance / Size of effect

P-value IS: Probability of data (or more extreme) given H0 is true: P(Data | H0)

Statistical significance != Practical significance

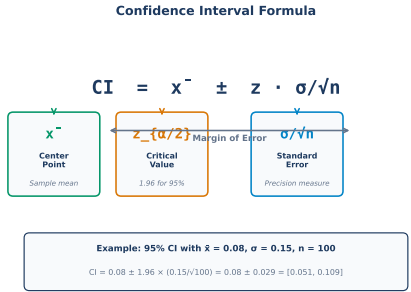
Self-study: p-value is NOT the probability that H0 is true

Self-Study: Confidence Intervals for Stock Returns



Self-study: 95% CI for expected portfolio return

Self-Study: Confidence Interval Formula



Self-study: CI = point estimate +/- margin of error

Self-Study: CI Formula Expanded

Confidence Interval Formula

```
CI = sample_mean +/- t_critical * (std / sqrt(n))
```

95% CI means:

If we repeated sampling 100 times,
~95 of the CIs would contain the true mean

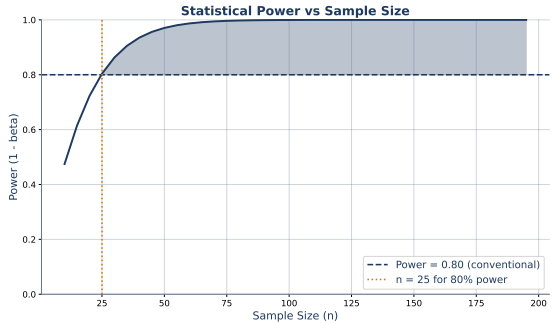
CI width depends on:
1. Confidence level (higher = wider)
2. Sample size (larger = narrower)
3. Variability (higher = wider)

CI vs p-value:

If 95% CI excludes H0 value,
then p-value < 0.05

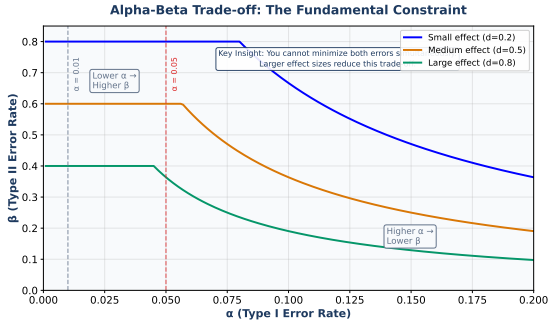
Self-study: larger n = narrower CI = more precision

Self-Study: Statistical Power



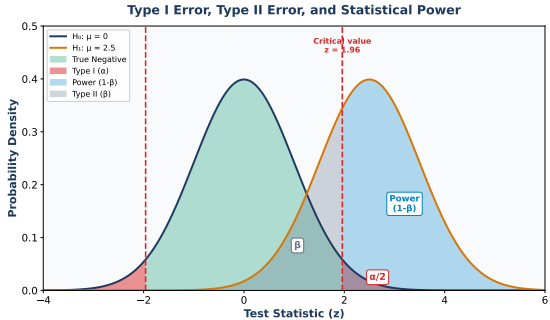
Self-study: power = 1 - beta = probability of detecting true effect

Self-Study: Alpha-Beta Trade-off



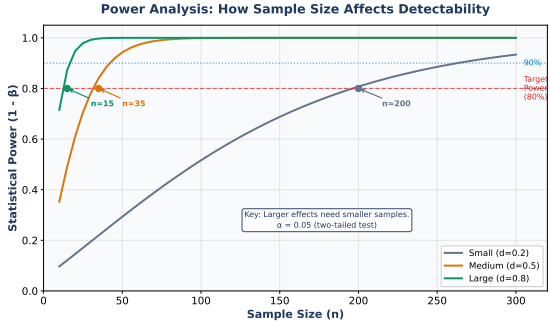
Self-study: reducing alpha increases beta and vice versa

Self-Study: Complete Rejection Regions



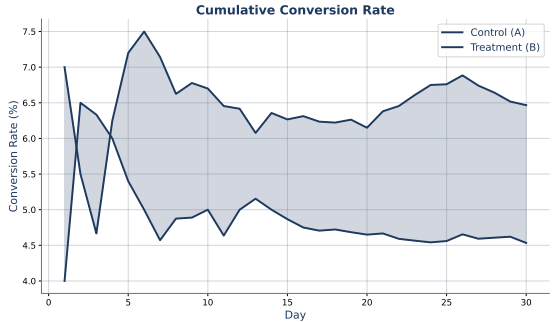
Self-study: one-tailed and two-tailed rejection regions compared

Self-Study: Power Curves



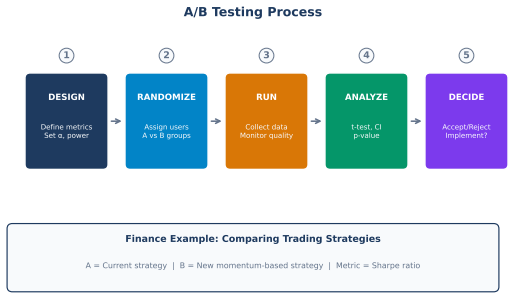
Self-study: how power changes with sample size and effect size

Self-Study: A/B Test Cumulative Performance



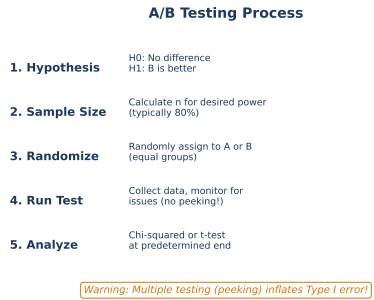
Self-study: track cumulative returns during A/B experiment

Self-Study: A/B Testing Process



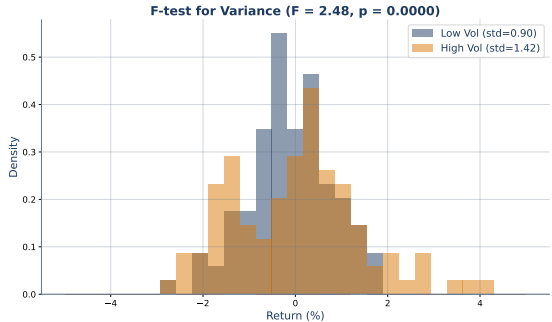
Self-study: design, run, analyze, conclude

Self-Study: A/B Testing Process Expanded



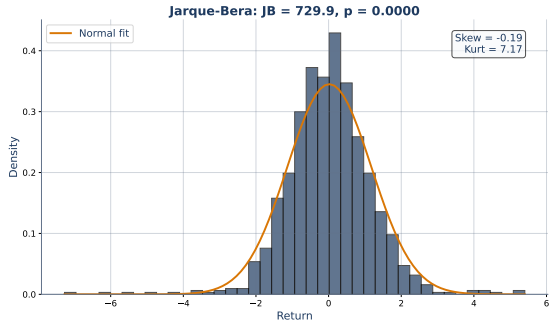
Self-study: complete A/B testing workflow

Self-Study: F-Test for Variance



Self-study: compare variances between portfolios or strategies

Self-Study: Jarque-Bera Normality Test



Self-study: test if returns follow normal distribution

Self-Study: Test Selection Guide

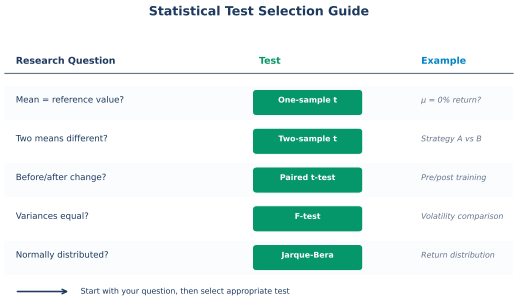
Common Finance Hypothesis Tests

t-test:	Test for alpha (excess returns)	`stats.ttest_1samp()`
F-test:	Compare variances (volatility)	`stats.f_oneway()`
Jarque-Bera:	Test for normality	`stats.jarque_bera()`
Ljung-Box:	Test for autocorrelation	`acorr_ljungbox()`
ADF Test:	Test for stationarity	`adfuller()`
ARCH Test:	Test for heteroskedasticity	`het_arch()`

Always check assumptions before applying any test!

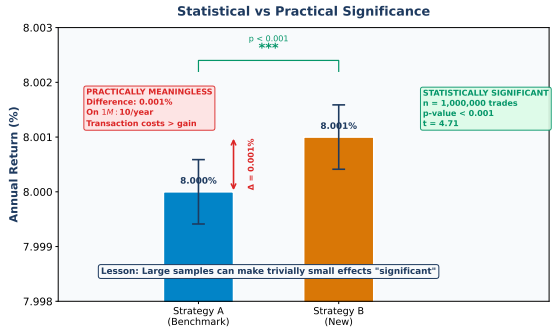
Self-study: choose appropriate test for your hypothesis

Self-Study: Test Selection Expanded



Self-study: t-test, F-test, Jarque-Bera, ADF compared

Self-Study: Practical vs Statistical Significance



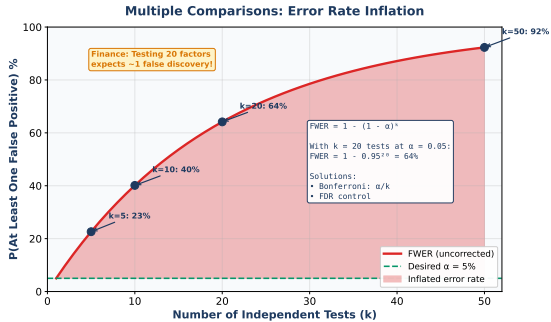
Cohen's d: $d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$

- $|d| < 0.2$: small $|d| \approx 0.5$: medium $|d| > 0.8$: large

---

Self-study: always report effect size alongside p-value

Self-Study: Multiple Comparisons Problem



Bonferroni: $\alpha_{adj} = \alpha/m$ where m = number of tests.
Testing 10 strategies at $\alpha = 0.05 \rightarrow$ use $\alpha = 0.005$ each.

Self-study: backtesting many strategies = data snooping = inflated results