

Lesson 15: Hypothesis Testing

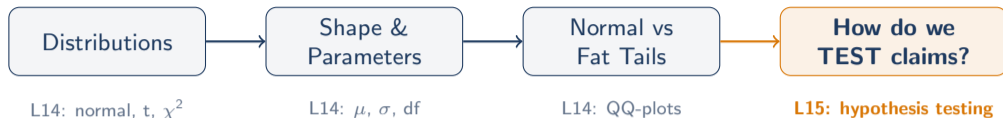
Data Science with Python – BSc Course

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

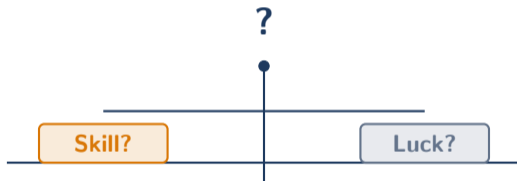
After this lesson, you will be able to:

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

The Central Challenge

A hedge fund claims 12% annual returns.
Is this skill or luck?



How much evidence do we need to decide?

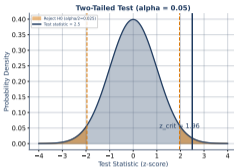
Hypothesis testing provides a principled framework for this exact question

The Logic of Hypothesis Testing

Hypothesis Testing: The Framework

Hypothesis Testing Process

- 1. State Hypotheses**
 H_0 : null hypothesis (status quo)
 H_1 : alternative (what we want to prove)
- 2. Choose Significance**
 $\alpha = 0.05$ (5% false positive rate)
- 3. Collect Data**
Sample from population
- 4. Calculate Test Statistic**
z, t, chi-squared, F, etc.
- 5. Make Decision**
Reject H_0 if p-value < α



Decision Rule

p-value < α
REJECT H_0
Evidence supports H_1

p-value $\geq \alpha$
FAIL TO REJECT H_0
Insufficient evidence

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

Note: "Fail to reject" is NOT the same as "accept H_0 "

- 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

Hypothesis Testing Process

1. State Hypotheses

H_0 : null hypothesis (status quo)
 H_1 : alternative (what we want to prove)

2. Choose Significance

$\alpha = 0.05$ (5% false positive rate)

3. Collect Data

Sample from population

4. Calculate Test Statistic

z, t, chi-squared, F, etc.

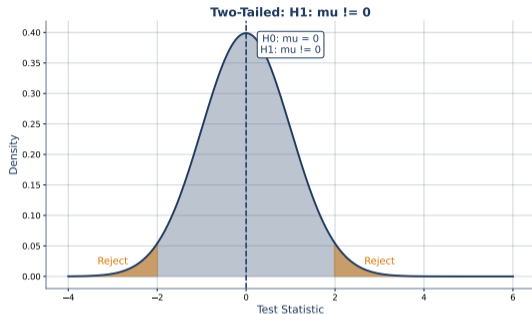
5. Make Decision

Reject H_0 if p-value < α

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 < α

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

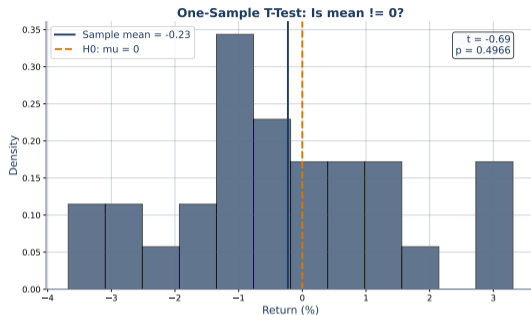
Rejection Regions: Two-Tailed Test



- $H_0: \mu = \mu_0$ 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



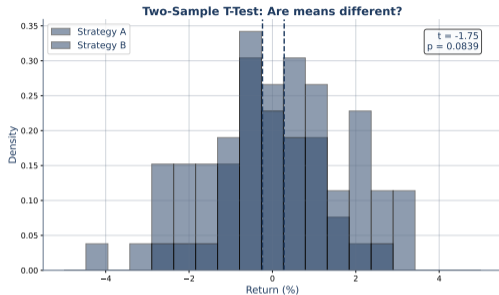
Formula: $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ $df = n - 1$

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)



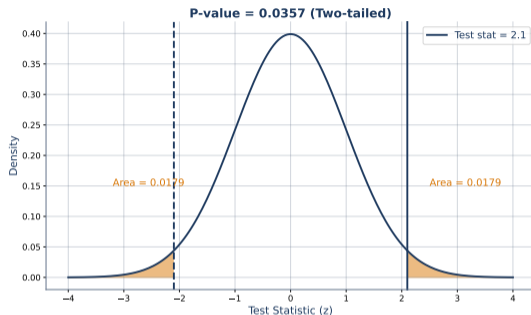
Formula:
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

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

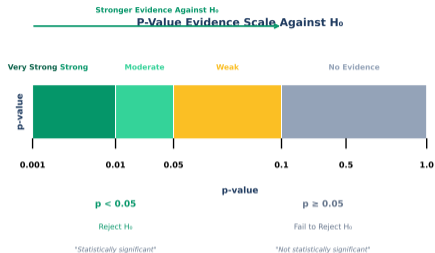
What Is a p-Value?



- **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

Checkpoint

Quick Check

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

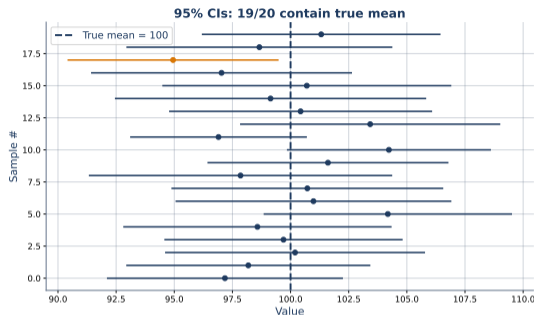
Answer

$$p = 0.03 < \alpha = 0.05 \Rightarrow \text{Reject } H_0$$

- 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

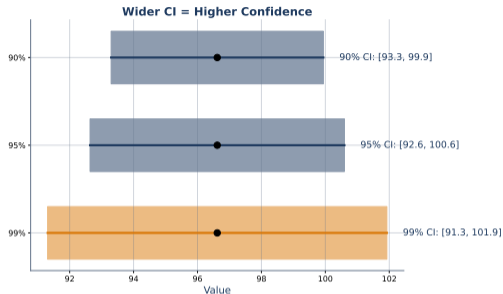


Formula: $CI = \bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$

- 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%

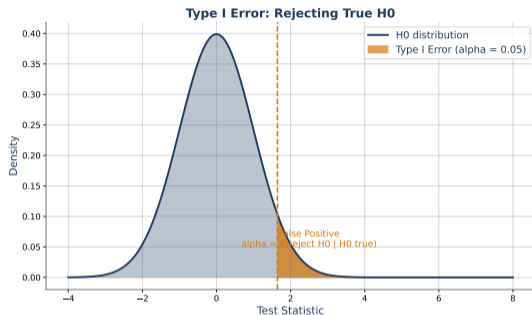


- Higher confidence → wider interval (more uncertainty)
- 95% is the standard in most research
- Trade-off: precision vs certainty

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

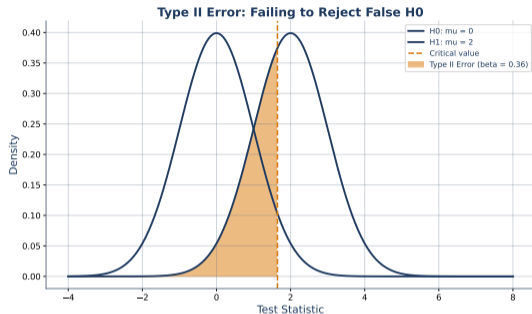
Type I Error: False Positive



- **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

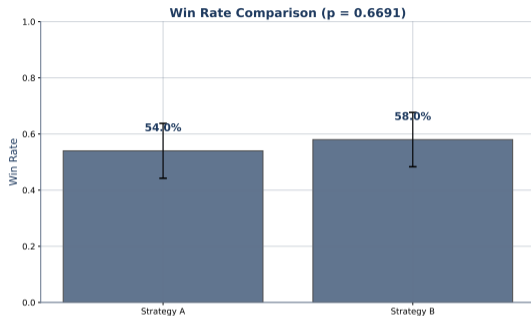
Decision Matrix

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

- 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

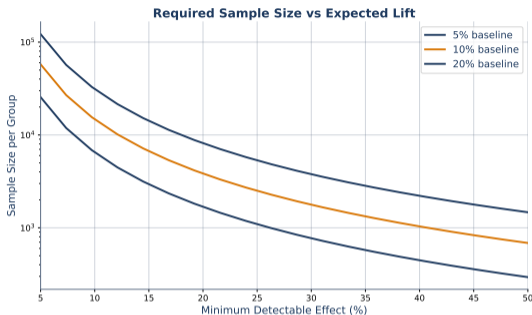
A/B Testing: Comparing Strategies



- 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?

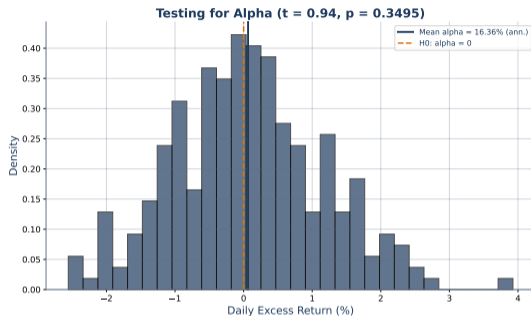


- More data → higher power → detect smaller effects
- Underpowered studies waste resources and miss real signals
- **Rule of thumb:** run a power analysis *before* collecting data

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:

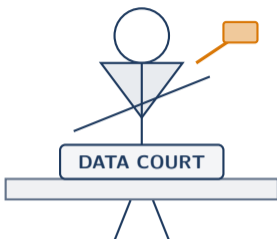
1. Load AAPL daily returns (use `yfinance` or sample data)
2. State hypotheses: $H_0: \mu = 0$ $H_1: \mu \neq 0$
3. Run one-sample t-test:

```
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:

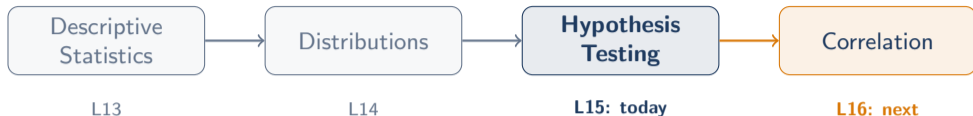
1. **Hypotheses come first:** formulate H_0 and H_1 before touching data
2. **p-value** = probability of data this extreme *if* H_0 is true (not the reverse!)
3. **t-tests** compare means: one-sample (vs value), two-sample (vs group)
4. **Two error types:** false positive (α) vs false negative (β) – 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