Hypothesis Testing

Foundations, Methods, and Examples

Yolymatics Tutorials
yolymatics007@gmail.com

Today's objectives

- Understand the logic of hypothesis testing and key terminology
- Choose and run common tests (means and proportions)
- Interpret p-values, significance, power, and practical significance
- Avoid common pitfalls and clearly communicate results

Big picture

What is hypothesis testing?

Idea

Use data to evaluate competing claims about a population parameter by comparing what we observed to what we'd expect if a baseline claim were true.

- **Null hypothesis** H_0: baseline/status quo (e.g., $\mu = 0$)
- **Alternative** H_1 or H_a: the effect/difference we care about
- **Test statistic**: summary of evidence (e.g., z, t)
- **p-value**: probability of results at least as extreme as observed if H_0 were true

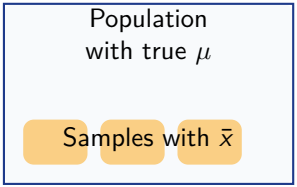
Parameters vs. statistics

Population parameter

- Unknown quantity (e.g., mean μ, proportion p)
- Fixed (does not vary across samples)

Sample statistic

- Computed from data (e.g., $\bar{x}$, $\hat{p}$)
- Random (varies across samples)



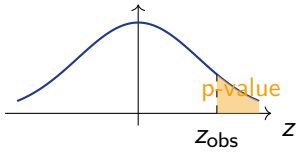
Errors and decisions

	Fail to reject H_0	**Reject H_0**
H_0 true	Correct	Type I error (α)
H_0 false	Type II error (β)	Correct (Power $= 1 - \beta$)

- **Significance level** α: tolerable Type I error rate (commonly 0.05)
- **Power**: probability we detect a true effect (larger is better)

p-values, visually

- p-value is computed assuming H_0 is true
- It is *not* the probability H_0 is true
- Smaller p-value = stronger evidence against H_0



Shaded area = p-value (one-sided)

Common tests

When to use which test?

Decision tree for choosing tests

Step 1: What parameter are you testing?

- Mean(s) $\rightarrow$ Use z-test or t-test
- Proportion(s) $\rightarrow$ Use z-test for proportions

Step 2: How many groups?

- One group $\rightarrow$ One-sample test
- Two groups $\rightarrow$ Two-sample or paired test

Step 3: Are observations paired/matched?

- Yes (same subjects, before/after) $\rightarrow$ Paired test
- No (independent groups) $\rightarrow$ Independent samples test

z-test vs t-test: When to use which?

Use z-test when:

- Testing **means** AND population σ is **known**
- Testing **proportions** with large n
- Large sample ($n \geq 30$) AND σ known

Use t-test when:

- Testing **means** AND population σ is **unknown**
- Small or large sample (any n)
- Use sample SD (s) to estimate σ

Most common situation

In practice, σ is usually unknown $\rightarrow$ **Use t-test for means**

Distribution requirements

For means (z-test or t-test)

Requirements:

- Data approximately normal, OR
- Large sample ($n \geq 30$) — Central Limit Theorem applies
- If $n < 30$ with non-normal data, consider non-parametric tests

For proportions (z-test)

Requirements (normal approximation):

- $np_0 \geq 10$ and $n(1 - p_0) \geq 10$ (one-sample)
- $n_1\hat{p}_1 \geq 10$, $n_1(1 - \hat{p}_1) \geq 10$, and same for group 2 (two-sample)
- If counts too small, use exact binomial test instead

General recipe

1. State H_0 and H_a (one- or two-sided)
2. Choose test and check assumptions
3. Compute test statistic
4. Get p-value (or critical value)
5. Conclude in context; consider effect size and CIs

Step-by-step: Complete hypothesis testing procedure

The 6 essential steps

Step 1: State hypotheses

- H_0: null hypothesis (status quo, usually "no effect" or specific value)
- H_a: alternative hypothesis (what you're trying to show)

Step 2: Choose significance level

- Common: $\alpha = 0.05$ or 0.01
- This is your tolerance for Type I error

Step 3: Check assumptions

- Independence, normality (or large n), appropriate counts

Step-by-step procedure (continued)

Steps 4-6

Step 4: Calculate test statistic

- Use appropriate formula (z, t, etc.)
- This measures how far your data is from H_0

Step 5: Find p-value

- Probability of seeing data this extreme if H_0 were true
- Use tables, technology, or formulas

Step 6: Make decision and conclude

- Compare p-value to α
- State conclusion in context of problem

Decision rules: Reject or fail to reject H_0?

Using p-value (most common)

- **If p-value** $\leq \alpha$: Reject H_0
  - Evidence *is* strong enough against H_0
  - Result is "statistically significant"
- **If p-value** $> \alpha$: Fail to reject H_0
  - Evidence *is not* strong enough against H_0
  - Result is "not statistically significant"

Important

We **never** "accept" H_0 — we only "fail to reject" it. Absence of evidence is not evidence of absence!

Decision rules: Using critical values (alternative)

Critical value method

Step 1: Find critical value(s) from tables based on α

Step 2: Compare test statistic to critical value(s)

- **Two-sided test:** Reject H_0 if $|\text{test stat}| > \text{critical value}$
- **Right-tailed:** Reject H_0 if $\text{test stat} > \text{critical value}$
- **Left-tailed:** Reject H_0 if $\text{test stat} < \text{-critical value}$

Example: For $\alpha = 0.05$, two-sided z-test

- Critical values: ± 1.96
- Reject H_0 if $|z| > 1.96$

One-sample z-test (mean, σ known)

Assumptions: independent sample; normal population or large n.

$$H_0: \ \mu = \mu_0, \qquad H_a: \ \mu \neq \mu_0 \ (\text{or } >, <)$$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \text{ under } H_0$$

p-value: one- or two-tailed area beyond $|z|$.

One-sample t-test (mean, σ unknown)

Assumptions: independent; approximately normal or large n.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \text{ under } H_0$$

Use when population SD is unknown (the usual case).

Two-sample t-test (independent groups)

$$H_0: \ \mu_1 - \mu_2 = 0, \qquad H_a: \ \mu_1 - \mu_2 \neq 0$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim t_\nu$$

with Welch–Satterthwaite df ν (robust to unequal variances).

Paired t-test (matched/paired data)

Convert to differences $d_i = x_{i,\text{after}} - x_{i,\text{before}}$ and test mean of d.

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} \sim t_{n-1}$$

One-proportion z-test

Large-sample test for a single proportion.

$$H_0 : \ p = p_0,$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$$

Rule of thumb: $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

Two-proportion z-test

Compare two independent proportions p_1 and p_2.

$$H_0 : \ p_1 - p_2 = 0, \qquad z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$ is the pooled proportion under H_0.

Design, power, and sample size

One- vs two-sided alternatives

- Use one-sided only when effects in the other direction are *impossible or irrelevant a priori*
- Two-sided is more conservative and standard in most analyses

Assumptions checklist

- Independence (study design, random sampling/assignment)
- Approximate normality for means (or large n via CLT)
- Sufficient counts for proportions
- No severe outliers/heteroscedasticity for t-tests

Power in plain terms

Power

Probability your test detects a true effect of a given size. Improves with larger effects, larger n, lower variability, and higher α.

- Report effect sizes (e.g., Cohen's d) and confidence intervals
- Consider practical vs. statistical significance

Back-of-envelope sample size

For mean with known σ (approximate):

$$n \approx \left(\frac{z_{1-\alpha/2}\,\sigma}{\mathrm{ME}} \right)^2$$

For proportion (worst-case $p = 0.5$):

$$n \approx \frac{z_{1-\alpha/2}^2\, 0.25}{\mathrm{ME}^2}$$

where ME is desired half-width of CI.

Practice Problems

Problem 1: One-sample t-test

Problem

A sample of $n = 25$ students has average score $\bar{x} = 78.4$ with $s = 10.2$. Test if the population mean differs from 75 at $\alpha = 0.05$.

Workspace:

Problem 1: Workspace (continued)

Problem 2: One-sample z-test

Problem

A factory claims their bolts have a mean diameter of 5.0 mm with known $\sigma = 0.12$ mm. A random sample of 40 bolts has $\bar{x} = 5.04$ mm. Test at $\alpha = 0.01$ if the true mean differs from 5.0 mm.

Workspace:

Problem 2: Workspace (continued)

Problem 3: One-sided t-test

Problem

A new teaching method is tested with 18 students. Their mean score is 82.5 with $s = 9.8$. The traditional method has a mean of 78. Test if the new method is better at $\alpha = 0.05$.

Workspace:

Problem 3: Workspace (continued)

Problem 4: Two-sample t-test

Problem

Group 1: $n_1 = 30$, $\bar{x}_1 = 72$, $s_1 = 8$
Group 2: $n_2 = 28$, $\bar{x}_2 = 68$, $s_2 = 9$
Test if the means differ at $\alpha = 0.05$.

Workspace:

Problem 4: Workspace (continued)

Problem 5: Two-sample comparison

Problem

Two brands of batteries are tested. Brand A ($n = 35$): $\bar{x} = 42.3$ hours, $s = 5.2$ hours. Brand B ($n = 40$): $\bar{x} = 39.8$ hours, $s = 6.1$ hours. Is there evidence Brand A lasts longer? Use $\alpha = 0.05$.

Workspace:

Problem 5: Workspace (continued)

Problem 6: Paired t-test

Problem

10 patients' blood pressure before and after treatment:
Differences (After - Before): $-8, -12, -5, -9, -15, -7, -11, -6, -10, -13$
Test if treatment reduces blood pressure at $\alpha = 0.05$.

Workspace:

Problem 6: Workspace (continued)

Problem 7: Paired data analysis

Problem

12 students take a test before and after tutoring. The mean difference is $\bar{d} = 7.5$ points with $s_d = 4.2$. Test if tutoring improves scores at $\alpha = 0.01$.

Workspace:

Problem 7: Workspace (continued)

Problem 8: One-proportion z-test

Problem

In $n = 200$ trials, $x = 118$ successes ($\hat{p} = 0.59$). Test $H_0 : p = 0.5$ vs $H_a : p \neq 0.5$ at $\alpha = 0.05$.

Workspace:

Problem 8: Workspace (continued)

Problem 9: Proportion test (one-sided)

Problem

A company claims that more than 30% of customers prefer their product. In a survey of 150 customers, 54 prefer it. Test the claim at $\alpha = 0.05$.

Workspace:

Problem 9: Workspace (continued)

Problem 10: Two-proportion z-test

Problem

Treatment A: 45 successes out of 100 trials
Treatment B: 38 successes out of 90 trials
Test if the success rates differ at $\alpha = 0.05$.

Workspace:

Problem 10: Workspace (continued)

Problem 11: Comparing proportions

Problem

Male voters: 132 out of 200 support a policy
Female voters: 145 out of 220 support the policy
Is there evidence of a gender difference in support? Use $\alpha = 0.01$.

Workspace:

Problem 11: Workspace (continued)

Problem 12: P-value interpretation

Problem

A hypothesis test yields $p = 0.032$.

1. What does this p-value mean in context?
2. What decision would you make at $\alpha = 0.05$?
3. What decision would you make at $\alpha = 0.01$?
4. Does this prove the null hypothesis is false?

Workspace:

Problem 12: Workspace (continued)

Problem 13: Type I and II errors

Problem

A drug company tests if a new drug reduces cholesterol.

1. State the null and alternative hypotheses
2. Describe a Type I error in context
3. Describe a Type II error in context
4. Which error is more serious? Explain.

Workspace:

Problem 13: Workspace (continued)

Problem 14: Assumptions check

Problem

You want to perform a one-sample t-test on a sample of $n = 12$ data points. The data shows strong right skewness and contains one extreme outlier.

1. Are the t-test assumptions met?
2. What could you do instead?
3. If you had $n = 100$ with the same skewness, would that change your answer?

Workspace:

Problem 14: Workspace (continued)

Problem 15: Sample size calculation

Problem

You want to estimate a population proportion with margin of error ± 0.04 at 95% confidence. Calculate the required sample size assuming:

1. Worst-case scenario ($p = 0.5$)
2. Previous estimate suggests $p = 0.3$

Workspace:

Problem 15: Workspace (continued)

Problem 16: Power concepts

Problem

1. If you decrease α from 0.05 to 0.01, what happens to power?
2. If you increase sample size, what happens to power?
3. If the true effect size is larger, what happens to power?
4. A test has power $= 0.85$. What does this mean?

Workspace:

Problem 16: Workspace (continued)

Problem 17: Complete analysis (means)

Problem

A researcher claims a new diet reduces weight. 20 people follow the diet for 8 weeks. Weight loss (kg): Mean $= 3.8$, SD $= 2.1$.

1. State appropriate hypotheses
2. Check assumptions
3. Compute test statistic
4. Find p-value and conclude at $\alpha = 0.05$
5. Calculate a 95% confidence interval

Workspace:

Problem 17: Workspace (continued)

Problem 17: Workspace (continued 2)

Problem 18: Complete analysis (proportions)

Problem

Before campaign: 120 out of 300 people support a policy (40%)
After campaign: 145 out of 280 people support it (51.8%)

1. State hypotheses for testing if support increased
2. Conduct appropriate test at $\alpha = 0.05$
3. Calculate effect size
4. Is the result practically significant?

Workspace:

Problem 18: Workspace (continued)

Problem 18: Workspace (continued 2)

Problem 19: Critical thinking

Problem

A study tests 50 different hypotheses and finds 3 significant results at $\alpha = 0.05$.

1. How many would you expect by chance alone?
2. What problem does this illustrate?
3. What should the researchers do?

Workspace:

Problem 19: Workspace (continued)

Problem 20: Design a study

Problem

You want to test if a new app improves math scores. Design a study including:

1. Null and alternative hypotheses
2. Type of test you'll use
3. Sample size justification
4. How you'll minimize bias
5. What you'll report beyond just p-values

Workspace:

Problem 20: Workspace (continued)

Problem 20: Workspace (continued 2)

Communication

Interpreting p-values (do and don't)

- Do: "If H_0 were true, we'd see results this extreme only about 1
- Don't: "The probability H_0 is true is 1
- Report effect size and CI; avoid dichotomizing by 0.05 alone

Common pitfalls

- P-hacking and multiple testing without correction
- Ignoring assumptions or study design
- Confusing absence of evidence with evidence of absence
- Over-reliance on arbitrary significance thresholds
- Ignoring effect sizes and practical significance

Thank you!

Yolymatics Tutorials

yolymatics007@gmail.com

Keep practicing and stay curious!