Introduction to Data Science

- Statistical Inference: Hypotesis testing and Significance -

Mauricio Molina

Keio University, Faculty of Economics

June 11, 2025

Contents

Hypotheses: Definition

Definition

A hypothesis test is a process that uses sample statistics to test a claim about the value of a population parameter.

- Two competing hypotheses:
  - H_a Alternative hypothesis: what the researcher wishes to support
  - H_0 Null hypothesis: the contradictory statement to H_a
- We begin by *assuming* H_0 is true.
- To support H_a, we must produce evidence that makes H_0 unlikely.
- Based on our sample data, we draw one of two conclusions:
  - **Reject** H_0 and conclude that H_a is true.
  - **Accept (do not reject)** H_0 as true.
- Note: "Accept H_0" does not prove it true, only that there is insufficient evidence against it.

Examples of Hypothesis Tests

Example 1: Electrician Wages in California

- Research question: Is the average hourly wage of electricians in California different from the national average of \$21?
- $H_a : \mu \neq 21, \quad H_0 : \mu = 21$
- Goal: Reject H_0 to conclude $\mu \neq 21$.

Example 2: Die Cutting Defect Rate

- A sheet-metal die cutting process currently produces 3% defectives.
- You wish to show that an adjustment decreases p, the defective proportion.
- $H_a : p < 0.03, \quad H_0 : p = 0.03$
- If you can reject H_0, conclude the adjusted process yields fewer than 3% defectives.

Two-Tailed vs One-Tailed Tests

- In Example 1, the alternative hypothesis is

$$H_a : \mu \neq 21,$$

which does not specify a direction (could be less *or* greater).
This is a two-tailed test.

- In Example 2, the alternative hypothesis is

$$H_a : p < 0.03,$$

which specifies a lower direction only.
This is a one-tailed (left-tailed) test.

Deciding with Sample Information

To decide whether to reject or accept H_0, we can use two pieces of information calculated from a sample, drawn from the population of interest:

- **Test statistic**:
  - A single number calculated from the sample.
  - Based on the best estimator for the parameter being tested.
- **p-value**:
  - The probability, assuming H_0 is true, of observing a test statistic as extreme as (or more extreme than) the one computed.

Example: Assessing Unusualness

- Test form: $H_0 : \mu = 21$ versus $H_a : \mu \neq 21$.
- Assume H_0 true ($\mu = 21$).
- Sample of $n = 100$ electricians:

$$\bar{X} = 22, \quad s = 2.$$

- Is $\bar{X} = 22$ unusual under H_0? We use two measures to decide.

Standardized Test Statistic

- Standard error of the mean:
$$\mathrm{SE} = \frac{s}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0.2.$$
- Test statistic:
$$z = \frac{\bar{X} - \mu_0}{\mathrm{SE}} = \frac{22 - 21}{0.2} = 5.$$
- $\bar{X} = 22$ lies 5 standard deviations above μ_0 — very unlikely if H_0 is true.

p-Value for Two-Tailed Test

What is the probability of observing $\bar{X} = 22$ or something even more unlikely if $\mu = 21$? The value $\bar{X} = 22$ lies 5 standard deviations above $\mu = 21$, but an equally "unlikely" value of would be one lying 5 standard deviations below $\mu = 21$.

- The p-value is
$$P(|Z| \geq 5) = P(Z > 5) + P(Z < -5) \approx 0 + 0 = 0.$$
- Such an outcome is essentially impossible under H_0.
- Conclusion: Strong evidence to reject H_0.

A Statistical Test of Hypothesis

Definition

The level of significance (significance level) α for a statistical test of hypothesis is

$$\alpha = P(\text{falsely rejecting } H_0) = P(\text{rejecting } H_0 \mid H_0 \text{ is true}).$$

This value α represents the maximum tolerable risk of incorrectly rejecting H_0. Once this significance level is fixed, the rejection region can be set to allow the researcher to reject H_0 with a fixed degree of confidence in the decision.

The Essentials of the Test

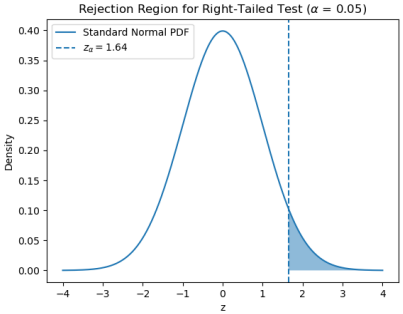
The sample mean $\bar{X}$ is the best estimate of the actual value of μ, which is presently in question. If H_0 is true and $\mu = \mu_0$, then $\bar{X}$ should be fairly close to μ_0. But if $\bar{X}$ is much larger than μ_0, this would indicate that $H_a : \mu > \mu_0$ might be true.

Since the sampling distribution of the sample mean $\bar{X}$ is approximately normal when n is large, the number of standard deviations that $\bar{X}$ lies from μ_0 can be measured using the test statistic

$$z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}},$$

which has an approximate standard normal distribution when H_0 is true and $\mu = \mu_0$.

The rejection region, shown in the figure, consists of values of z which are much larger than expected.



Since the significance level α is defined as the probability of rejecting H_0 when it is true, it is the area under the curve above the rejection region—the shaded area in the figure. The critical value of z cutting off area α in the right tail is called z_α.

Example 4: Weekly Earnings of Social Workers The average weekly earnings for female social workers is \$670. Do men in the same positions have average weekly earnings higher than \$670? A random sample of $n = 40$ male social workers showed $\bar{X} = 725$ and $s = 102$. Test the hypotheses using $\alpha = 0.01$.

Hypotheses & Test Statistic

You would like to show that the average weekly earnings for men are higher than \$670, the women's average. Hence, if μ is the average weekly earnings for male social workers, you can set out the formal test of hypothesis in steps:

Null and alternative hypotheses:

$$H_0 : \mu = 670 \quad \text{vs.} \quad H_a : \mu > 670$$

Test statistic: Using the sample information, with s as an estimate of the population standard deviation, calculate

$$z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{725 - 670}{102/\sqrt{40}} \approx 3.41$$

Rejection Region & Critical Value

- One-tailed (right-tailed) test at $\alpha = 0.01$.
- Critical value $z_{0.01}$ satisfies

$$P(Z > z_{0.01}) = 0.01 \quad \Longrightarrow \quad z_{0.01} \approx 2.33.$$

- Rejection region: $z > 2.33$.

Conclusion

$$z_{\text{obs}} = 3.41 \ > \ z_{0.01} = 2.33$$

Therefore, z_{obs} lies in the rejection region. Reject H_0. Conclude that male social workers earn on average more than \$670 weekly. The probability that you made an incorrect decision is $\alpha = 0.01$.

Python Example

Python code

```
import numpy as np
from scipy.stats import norm

# Given data
xbar, mu0, s, n = 725, 670, 102, 40
alpha = 0.01

# Compute test statistic and critical value
z_obs    = (xbar - mu0) / (s / np.sqrt(n))
z_crit   = norm.ppf(1 - alpha)

# Decision
if z_obs > z_crit:
 decision = "Reject H0"
else:
 decision = "Fail to reject H0"

print(f"z_obs = {z_obs:.2f}, z_crit = {z_crit:.2f}")
print(decision)
```

Large-Sample Statistical Test for μ

1. Null hypothesis:

$$H_0 : \mu = \mu_0$$

2. Alternative hypothesis:

$$\begin{cases} H_a : \mu > \mu_0 \quad (\text{or } H_a : \mu < \mu_0) \quad (\text{one-tailed}) \\ H_a : \mu \neq \mu_0 \qquad (\text{two-tailed}) \end{cases}$$

3. Test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \quad (\text{or estimated as}) \quad z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

Rejection Region: Reject H_0 When...

One-Tailed Test

$z > z_\alpha$ (or $z < -z_\alpha$ when $H_a : \mu < \mu_0$)

Assumptions:

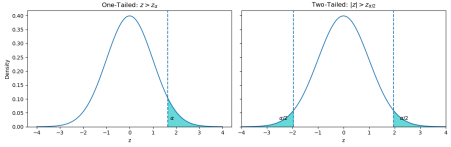
The n observations in the sample are randomly selected from the population and n is large (say, $n \geq 30$).

Rejection regions:

- *Right-tailed:* $z > z_\alpha$
- *Two-tailed:* $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

Two-Tailed Test

$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$



Calculating the p-Value

In the previous examples, different significance levels may lead to different conclusions. To avoid ambiguity, some experimenters prefer a variable level of significance called the p-value.

Definition

The *p-value* or observed significance level of a test is the smallest value of α for which H_0 can be rejected. It is the *actual risk* of committing a Type I error if H_0 is rejected based on the observed test statistic. The p-value measures the strength of the evidence against H_0.

- *Small p-values* indicates the observed test statistic lies far from the hypothesized value of μ. This presents strong evidence that H_0 is false and should be rejected.
- *Large p-values* indicate the observed statistic is not far from the hypothesized mean and do not support rejecting H_0.

Definition

If the p-value $\leq$ a preassigned significance level α, then H_0 can be rejected, and the results are said to be *statistically significant* at level α.

Example 5 A quality-control manager wants to know whether the daily yield at a chemical plant—which has averaged 880 tons for years—has changed in recent months.
A random sample of $n = 50$ days gives

$$\bar{X} = 871, \quad s = 21.$$

Calculate the p-value for this two-tailed test and draw conclusions.

Solution steps:

$$H_0 : \mu = 880 \quad \text{vs.} \quad H_a : \mu \neq 880$$

Sampling distribution (large n):

$$z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{871 - 880}{21/\sqrt{50}} \approx -3.03.$$

Rejection region (two-tailed): $|z| > z_{\alpha/2}$. The p-value is

$$P(|Z| \geq 3.03) = P(Z > 3.03) + P(Z < -3.03) = (1 - 0.9988) + 0.0012 = 0.0024.$$

Since 0.0024 is less than 0.01 (and 0.05), reject H_0.

Python Example: Computing z and the p-Value

Python code

```
# Given data
xbar, mu0, s, n = 871, 880, 21, 50

# Compute test statistic
z_obs = (xbar - mu0) / (s / np.sqrt(n))

# Two-tailed p-value  sf = 1-cdf
p_val = 2 * norm.sf(abs(z_obs))

print(f"Observed z = {z_obs:.2f}")
print(f"Two-tailed p-value = {p_val:.4f}")

# Conclusion at alpha = 0.01
alpha = 0.01
if p_val <= alpha:
 print("Reject H0: evidence of change in daily yield.")
else:
 print("Fail to reject H0: no evidence of change.")
```

Classifying p-Values: A "Sliding Scale"

- If the p-value is less than 0.01, reject H_0. The results are **highly significant**.
- If the p-value is between 0.01 and 0.05, reject H_0. The results are **statistically significant**.
- If the p-value is between 0.05 and 0.10, you usually do *not* reject H_0. The results are only **tending toward statistical significance**.
- If the p-value is greater than 0.10, do not reject H_0. The results are **not statistically significant**.

Example 6: Sodium Intake Test Standards recommend that Japanese should not exceed an average daily sodium intake of 3300 mg. A sample of $n = 100$ Japanese yields:

$$\bar{X} = 3400 \text{ mg}, \quad s = 1100 \text{ mg}.$$

Conduct a one-tailed test at $\alpha = 0.05$ to see if the mean exceeds 3300 mg.

$$H_0 : \mu = 3300 \quad \text{versus} \quad H_a : \mu > 3300$$

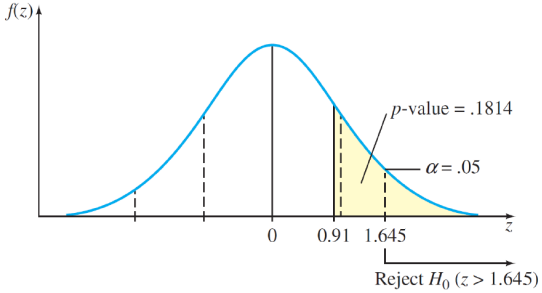
$$z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{3400 - 3300}{1100/\sqrt{100}} = 0.91$$

Critical Value Approach

The critical-value approach (one-tailed, $\alpha = 0.05$):

$$z_\alpha = 1.645.$$

Rejection region: $z > 1.645$.



Since $z_{\text{obs}} = 0.91$ is not greater than 1.645, we *do not* reject H_0.

p-Value Approach

Calculate the p-value for $z = 0.91$:

$$p\text{-value} = P(Z > 0.91) = 1 - 0.8186 = 0.1814.$$

Reject H_0 only if p-value ≤ 0.05.
Here, $0.1814 > 0.05$, so we *do not* reject H_0.
Conclusion: There is not enough evidence to indicate that the average daily sodium intake exceeds 3300 mg.

Advantages of the p-Value Approach

- Statistical software typically reports the p-value directly. In python (see `ttest_rel` from `scipy.stats`)
- You can evaluate your results using *any* significance level you choose. Many researchers report the smallest α for which their results are statistically significant.

Python Example: Sodium Intake Test

Python code

```
import numpy as np
from scipy.stats import norm

# Given data
xbar, mu0, s, n = 3400, 3300, 1100, 100
alpha = 0.05

# Compute test statistic and p-value
z_obs = (xbar - mu0) / (s / np.sqrt(n))
p_val = 1 - norm.cdf(z_obs)

print(f"Observed z = {z_obs:.2f}")
print(f"One-tailed p-value = {p_val:.4f}")

# Decision
if p_val <= alpha:
 print("Reject H0: mean sodium > 3300 mg")
else:
 print("Fail to reject H0: no evidence mean > 3300 mg")
```

Two Types of Errors (1 of 2)

There are two possible errors in a statistical test.

1. The researcher might reject H_0 when it is really true.
2. The researcher might accept H_0 when it is really false.

For a statistical test, these two types of errors are defined as Type I and Type II errors, shown in the decision table below.

	Null Hypothesis	
Decision	**True**	**False**
Reject H_0	Type I Error (α)	Correct
Accept H_0	Correct	Type II Error (β)

Two Types of Errors (2 of 2)

Definition

A **Type I error** for a statistical test happens if you reject the null hypothesis when it is true. The probability of making a Type I error is denoted by the symbol α.

Definition

A **Type II error** for a statistical test happens if you accept the null hypothesis when it is false and some alternative hypothesis is true. The probability of making a Type II error is denoted by the symbol β.

The Power of a Statistical Test

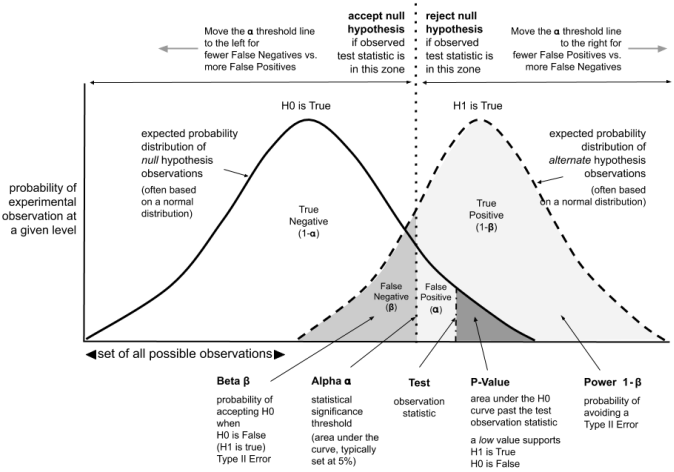
Definition

The *power* of a statistical test, given as

$$1 - \beta \;=\; P(\text{reject } H_0 \mid H_a \text{ is true}),$$

measures the ability of the test to detect an effect when the alternative hypothesis is true.

A graph of $1 - \beta$, the probability of rejecting H_0 when in fact H_0 is false, as a function of the true value of the parameter of interest is called the *power curve* for the test. Ideally, you would like α to be small and the power $(1 - \beta)$ to be large.



Statistical Power. Source: The Science of Machine Learning and AI; https://www.ml-science.com/statistical-power-of-a-test

Large-Sample Test for the Difference Between Two Means

The statistic summarizing sample information for $\mu_1 - \mu_2$ is the difference in sample means $\bar{x}_1 - \bar{x}_2$. Its true standard error is

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{(estimated by)} \quad \mathrm{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

We form the z statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\mathrm{SE}} \quad ,$$

which (for large n_1, n_2) is approximately standard normal under H_0.

Assumptions: Samples are random and independent from two populations, each with $n_i \geq 30$.

Procedure:

1. **Null hypothesis:** $H_0 : (\mu_1 - \mu_2) = D_0$. Often $D_0 = 0$.
2. **Alternative hypothesis:**

$$\begin{cases} H_a : (\mu_1 - \mu_2) > D_0 \quad (\text{or } < D_0) & \text{one-tailed,} \\ H_a : (\mu_1 - \mu_2) \neq D_0 & \text{two-tailed.} \end{cases}$$

3. **Test statistic:**

$$z \approx \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Example: To determine whether car ownership affects a student's academic achievement, random samples of 100 car owners and 100 nonowners were drawn from the student body. The grade point average for the $n_1 = 100$ nonowners had an average and variance equal to $\bar{x}_1 = 2.70$, $s_1^2 = 0.36$, and for the $n_2 = 100$ car owners: $\bar{x}_2 = 2.54$, $s_2^2 = 0.40$. Do the data present sufficient evidence to indicate a difference in the mean achievements between car owners and nonowners? Test at $\alpha = 0.05$ whether $\mu_1 - \mu_2 \neq 0$.

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_a : \mu_1 - \mu_2 \neq 0$$

$$z = \frac{(2.70 - 2.54) - 0}{\sqrt{\frac{0.36}{100} + \frac{0.40}{100}}} = 1.84.$$

Critical-value approach:

Two-tailed $\alpha = 0.05$ gives $z_{0.025} = 1.96$. Since $|1.84| < 1.96$, do not reject H_0.

p-value approach:

$$p\text{-value} = P(Z > 1.84) + P(Z < -1.84) = (1 - 0.9671) + 0.0329 = 0.0658.$$

Since $0.05 < 0.0658 < 0.10$, we *cannot* reject H_0 at $\alpha = 0.05$ (but could at 0.10).

Python code:

```
# Data
x1, s1, n1 = 2.70, np.sqrt(0.36), 100
x2, s2, n2 = 2.54, np.sqrt(0.40), 100

# Compute z and p-value
num    = (x1 - x2)
den    = np.sqrt(s1**2/n1 + s2**2/n2)
z_obs  = num / den
p_val  = 2 * norm.sf(abs(z_obs))

print(f"z = {z_obs:.2f}, p-value = {p_val:.4f}")
```

Assignment 9

Answer all three questions in a Jupyter Notebook. Show your Python code (when requested) and a short explanation for every result. Upload the completed `.ipynb` to K-LMS by next Tuesday at midnight.

Q1. From the Student Performance dataset:

- Load the two student datasets: `student-mat.csv` and `student-por.csv`.
- Merge them on the common identifiers (e.g. `school, sex, age, address, famsize, Pstatus, guardian`, etc.).
- Compute and report:

$$\bar{G}_{1,\text{math}} = \text{mean of G1 in math}, \quad \bar{G}_{1,\text{por}} = \text{mean of G1 in Portuguese}.$$

Q2. Test whether the mean G_1 is the same in math and Portuguese.

- Formulate hypotheses:

$$H_0 : \mu_{\text{math}} = \mu_{\text{por}} \quad \text{vs.} \quad H_a : \mu_{\text{math}} \neq \mu_{\text{por}}$$

- Use a two-sample (or paired) large-sample z and test for $\alpha = 5\%$ and $\alpha = 1\%$.
- Compute the corresponding p-value
- Make the interpretion based on results of the tests.

Q3. Consider the difference in `studytime` between courses

- State hypotheses for $\mu_{\text{studytime, math}}$ vs. $\mu_{\text{studytime, por}}$.
- Perform an appropriate test at $\alpha = 0.05$.
- Compute the p-value and state your conclusion: *Is there evidence of a difference in study time between math and Portuguese?*