Chapter 2: Hypothesis Testing

A Statistical Approach to Decision Making

Prof. Alok Shukla
with help from Gemini and Chat GPT

Department of Physics
IIT Bombay, Powai, Mumbai 400076

Course Name: AI and Data Science (PH 227):
Part 1 (Data Science)

What is a Hypothesis?

- In plane English, a hypothesis is a proposed explanation for a phenomenon.

Definition

Hypothesis testing is a statistical procedure to make inferences about a population parameter using sample data.

- In statistics, hypothesis testing (HT) starts with a statement about a population parameter (e.g., mean, proportion, standard deviation).
- **Example:** "The average height of male students is 175 cm."

Hypothesis Testing

HT Procedure (In short)

- Start with a claim (hypothesis).
- Collect sample data.
- Use probability theory to decide whether to accept or reject the claim.

The Two Types of Hypotheses

- **1. Null Hypothesis (H_0):**
  - This is the statement of "no effect," "no difference," or "no relationship."
  - It's the assumption we start with and try to find evidence against.
  - **Example:** H_0: The average height of male students is equal to 175 cm ($\mu = 175$).
- **2. Alternative Hypothesis (H_a or H_1):**
  - This is the statement we hope to find evidence for.
  - It contradicts the null hypothesis.
  - **Example:** H_a: The average height of male students is not equal to 175 cm ($\mu \neq 175$).

The Core Idea of Hypothesis Testing

- We collect a sample and calculate a **test statistic** (z-score, t-score, χ^2, etc.).
- This statistic measures how far our sample result is from the null hypothesis's claim.
- We then determine the **p-value**, which is the probability of observing a test statistic as extreme as, or more extreme than, the one we got, assuming the null hypothesis is true.
- **Key Concept:** A small p-value suggests that our observed data is unlikely if the null hypothesis is true, leading us to **reject the null hypothesis**.

The Steps of Hypothesis Testing

1. **State the Hypotheses:** Clearly define the null (H_0) and alternative (H_a) hypotheses.
2. **Set the Significance Level (α):** This is the threshold for rejecting the null hypothesis. Common values are 0.05 (5%) or 0.01 (1%). If the p-value is less than or equal to α, we reject H_0.
3. **Choose the Correct Test:** Select the appropriate statistical test such as z-test, t-test, χ^2 test, based on the type of data and the question.
4. **Calculate the Test Statistic and p-value:** Use the sample data to compute the test statistic and find its corresponding p-value either from a table, or a software
5. **Make a Decision:** Compare the p-value to the significance level (α) and decide whether to reject or fail to reject the null hypothesis.
6. **State a Conclusion:** Interpret the decision in the context of the original problem.

Types of Errors That Can Occur in Hypothesis Testing

- **Type I Error (α):**
  - **Definition:** Rejecting the null hypothesis when it is actually true.
  - **Analogy:** A court convicts an innocent person.
  - The probability of a Type I error is equal to our significance level, α.
- **Type II Error (β):**
  - **Definition:** Failing to reject the null hypothesis when it is actually false.
  - **Analogy:** A court acquits a guilty person.
  - **Power of the test:** The probability of correctly rejecting a false null hypothesis is $1 - \beta$.

Which test to use?

- When performing hypothesis testing, the choice of the correct statistical test is crucial.
- The test you choose depends on:
  1. The type of data you have (e.g., numerical, categorical).
  2. The size of your sample.
  3. Whether you know the population standard deviation.
- We will explore three of the most common tests.

The Z-Test

- **When to Use:**
  - To compare a sample mean to a population mean.
  - Used when the sample size is large ($n \geq 30$).
  - **Crucially**, used when the **population standard deviation (σ)** is **known**.
- **Test Statistic Formula:**

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

  - $\bar{x}$ is the sample mean.
  - μ is the population mean.
  - σ is the population standard deviation.
  - n is the sample size.
- **Example:** Testing if a large sample of students from a school has a different average score than the national average, where the national standard deviation is known.

The Z distribution function

The Z distribution function is nothing but the normal distribution function with $\mu = 0$ and $\sigma = 1$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

for $z \in [-\infty, \infty]$.

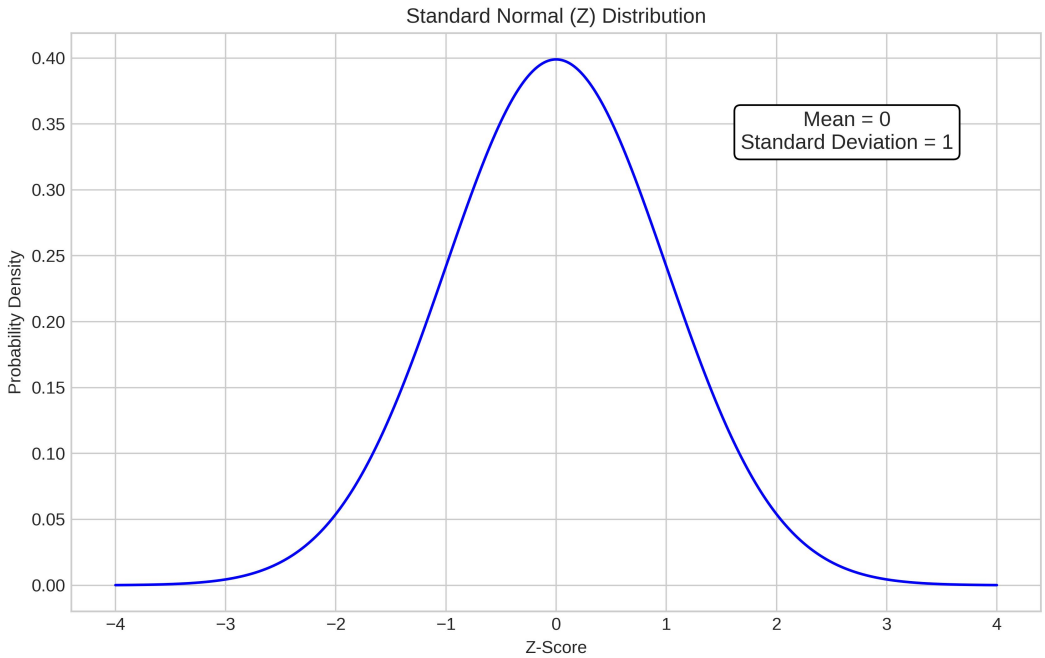

Figure: The z distribution function

The T-Test

- **When to Use:**
  - To compare a sample mean to a population mean.
  - Used when the sample size is small ($n < 30$).
  - **Crucially**, used when the **population standard deviation (σ)** is **unknown**. We use the sample standard deviation (s) instead.
- **Test Statistic Formula:**

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

  - s is the sample standard deviation.
  - This test uses a t-distribution, which accounts for the extra uncertainty from estimating the population standard deviation.
  - The shape of the t-distribution depends on the **degrees of freedom** ($df = n - 1$).
- **Example:** Testing if a new drug significantly changes blood pressure in a small group of patients, where the population standard deviation is not known.

The t-distribution function

The t-distribution function is known in the literature as "student's t-distribution", e.g., see wikipedia. The probability distribution function (or simply t-distribution) is given by the formula

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2},$$

where ν is the total # of degrees of freedom, and Γ is the gamma function.

- This distribution looks quite similar to the normal one, and is symmetric about the mean
- However, compared to the normal distribution, the t distribution has heavier tails as is obvious from the following figure

t-distribution function...

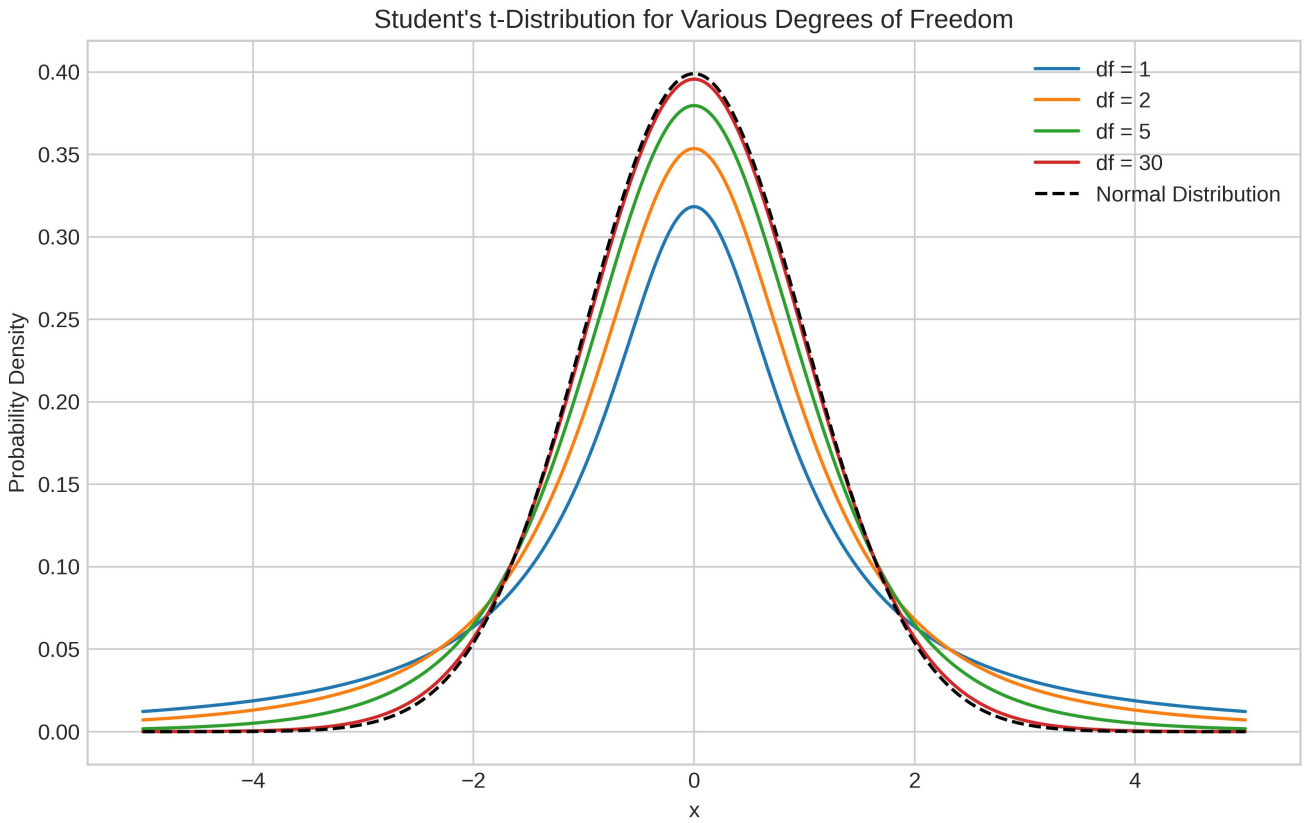

Figure: T distribution function for different values of degrees of freedom. For comparison, the normal distribution function with $\mu = 0$ and $\sigma = 1$ is also plotted.

The Chi-Squared Test (χ^2)

- **When to Use:**
  - To analyze **categorical data**.
  - It compares observed frequencies to expected frequencies.
- **Two Primary Uses:**
  1. **Test of Independence:** To determine if there is a statistically significant relationship between two categorical variables (e.g., gender and political affiliation).
  2. **Goodness-of-Fit Test:** To see if a sample distribution matches a hypothetical or expected population distribution (e.g., do survey responses match a known demographic breakdown?).

Summary: Choosing the Right Test

Test	**Data Type**	**Sample Size**	**Key Condition**
Z-Test	Numerical	Large ($n \geq 30$)	Population σ is Known
T-Test	Numerical	Small ($n < 30$)	Population σ is Unknown
Chi-Squared	Categorical	Any	Compares Frequencies

- This table serves as a quick guide for selecting the appropriate statistical test.

Example: one (right)-tailed Test

- **Problem** A new fertilizer is tested to see if it increases the average crop yield, which is known to be 10 tons per acre with a population standard deviation of 1.5 tons. A sample of 30 acres using the new fertilizer has an average yield of 10.8 tons. At a 5% significance level, can we conclude that the new fertilizer increases the average yield?
- **Soln:**

$H_0 : \mu = 10$ average yield is 10 tons/acre

$H_1 : \mu > 10$ average yield is more than 10 tons/acre

Here $\mu = 10$, $\sigma = 1.5$, $n = 30$, $\bar{x} = 10.8$, $\alpha = 0.05$, therefore, the test statistic and the z-score will be

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{10.8 - 10}{1.5/\sqrt{30}} = 2.92$$

$$z(\alpha = 0.05) = 1.64$$

Because $z = 2.92 > z_{crit} = 1.64$, H_0 has to be rejected and H_1 is correct

Plot of the right-tailed test

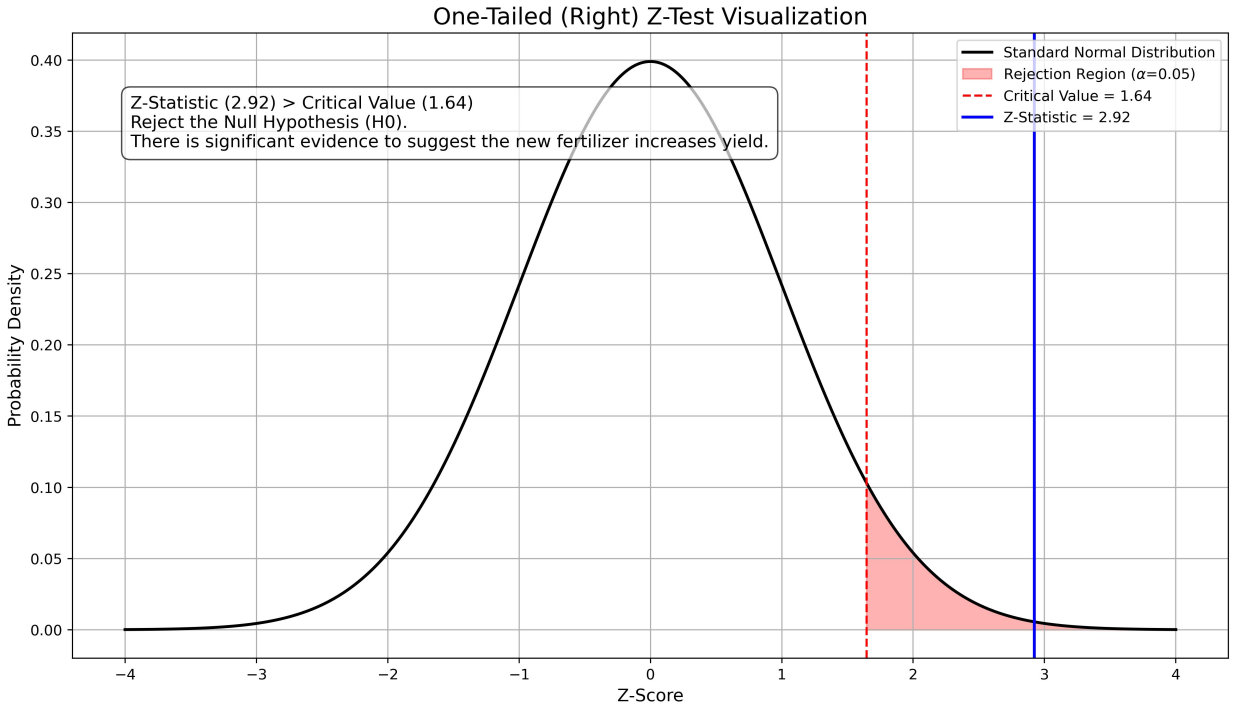

Figure: Here H_0 is that yield is not greater than the mean, while H_1 is that it is greater than mean

Example: Left-Tailed Test

Problem: A light bulb manufacturer claims their bulbs have a mean lifespan of at least 800 hours. The population standard deviation is known to be 60 hours. A sample of 40 bulbs is tested, and their average lifespan is found to be 785 hours. At a 5% significance level, is there enough evidence to conclude that the mean lifespan is less than 800 hours?
Soln: Here $\mu = 800$, $\sigma = 60$, $n = 40$, $\bar{x} = 785$, $\alpha = 0.05$, therefore, the test statistic and the z-score will be

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{785 - 800}{60/\sqrt{40}} = -1.581$$

$H_0 : \mu = 800$	average lifetime of bulbs is 800 hours
$H_1 : \mu < 800$	average lifetime of bulbs is less than 800 hours

Because $z = -1.58 > z_{crit} = -1.64$, we fail to reject H_0

Plot of the Left-Tailed Test

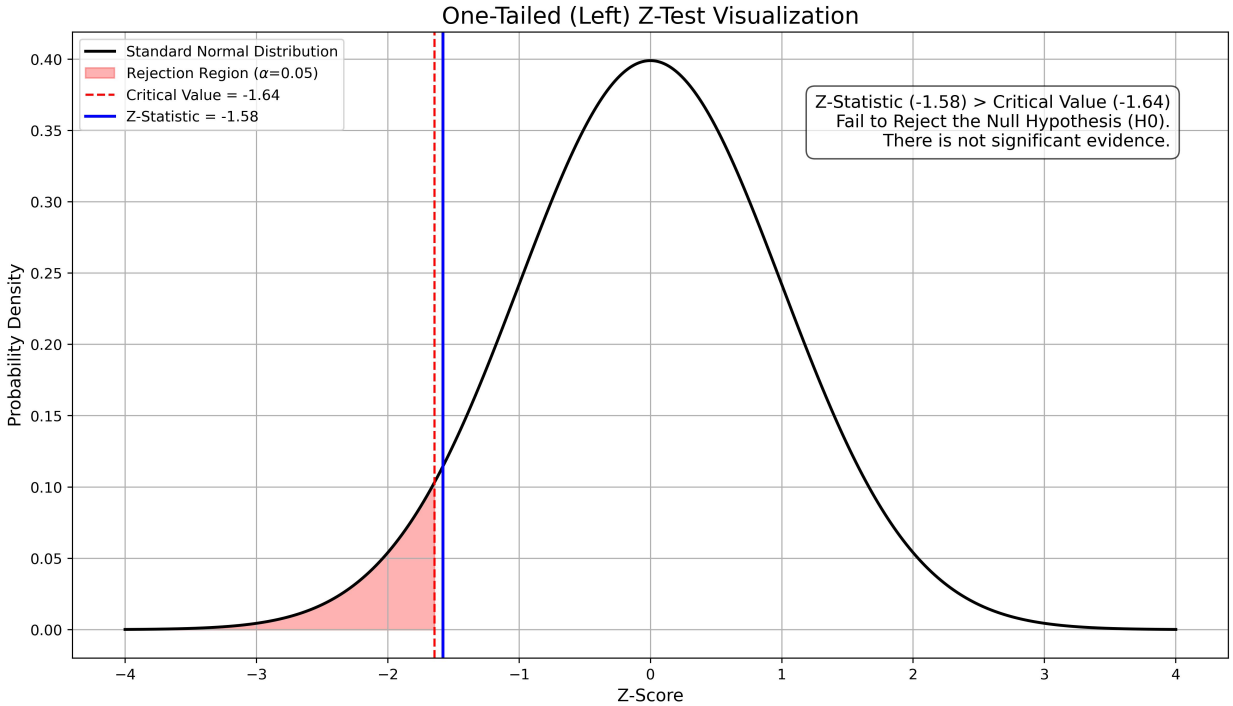

Figure: Here H_0 is that average liftetime of the bulbs is 800 hours, while H_1 is that it is less than that

Example: A two-tailed Z-test for the Mean

- Suppose we test:

$$H_0 : \mu = 100 \qquad H_1 : \mu > 100$$

  with $\sigma = 15$, $n = 36$, sample mean $\bar{x} = 104$, and $\alpha = 0.05$.

- Test statistic:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{104 - 100}{15/\sqrt{36}} = 1.6$$

- Critical value: $z_{0.05} = 1.645$
- Since $1.6 < 1.645$, **we fail to reject** H_0.

Z-test for the mean: the figure

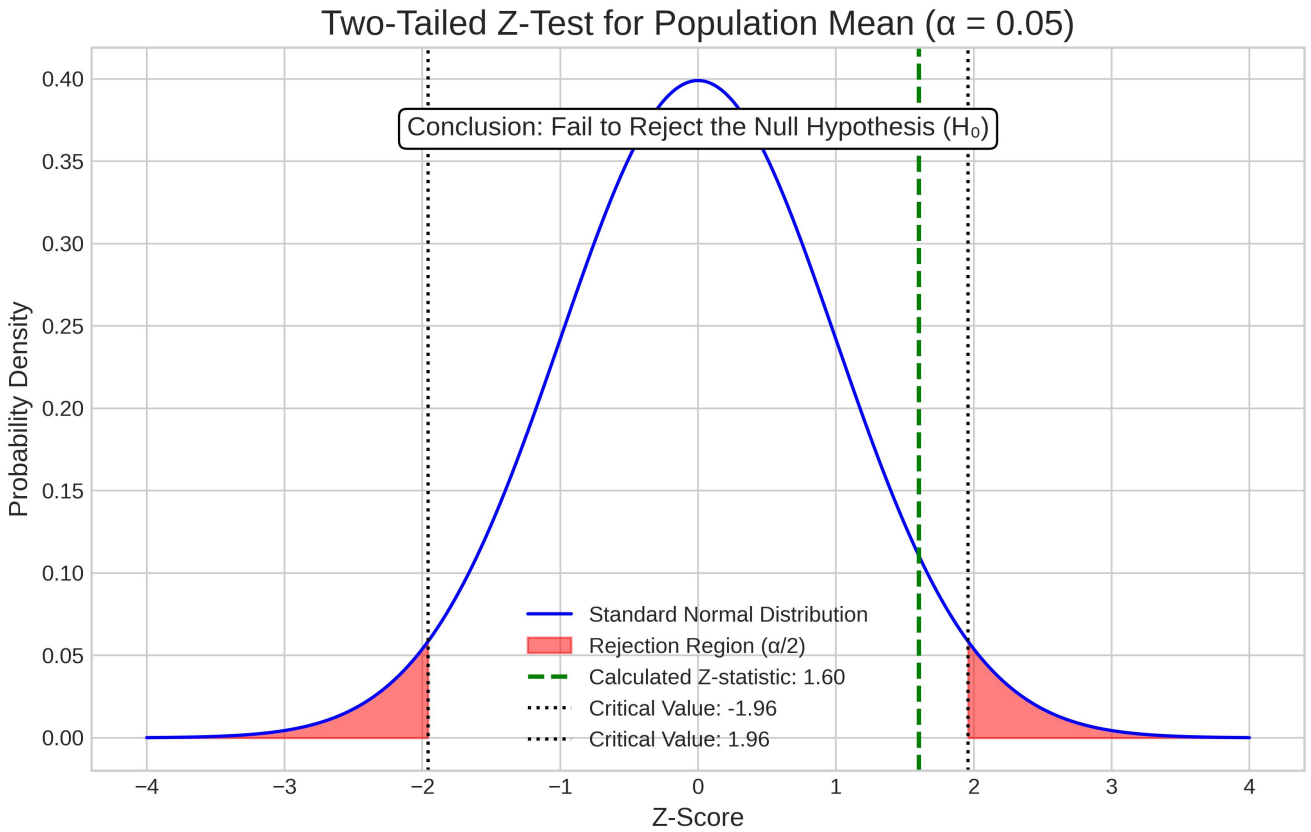

Figure: Here H_0 is that $\mu = 100$, and H_1 is that $\mu \neq 100$.Therefore, a two-tailed test is needed, which determines that there is no statistically significant evidence against H_0.

Example 2: Another two-tailed test for the mean

Problem: A coffee shop claims its average daily sales are \$1200. The manager believes this has changed. The population standard deviation of daily sales is known to be \$150. A random sample of 50 days is taken, and the average daily sales is found to be \$1150. At a 5% significance level, is there enough evidence to say the average daily sales are no longer \$1200?

Soln: Clearly

$$H_0 : \mu = 1200 \qquad H_1 : \mu \neq 1200$$

Here $\mu = 1200$, $\sigma = 150$, $n = 50$, $\bar{x} = 1150$, $\alpha = 0.05$, therefore, the test statistic and the z-score will be

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{1150 - 1200}{150/\sqrt{50}} = -2.357$$

Significance level: $\alpha = 0.05$, but $z_{crit} = z(\alpha/2) = -1.96$

- Critical values: ± 1.96 for z, therefore we reject H_0 because $|z| > 1.96$
- Therefore, the daily sales have indeed changed

Plot of the Two-Tailed Test

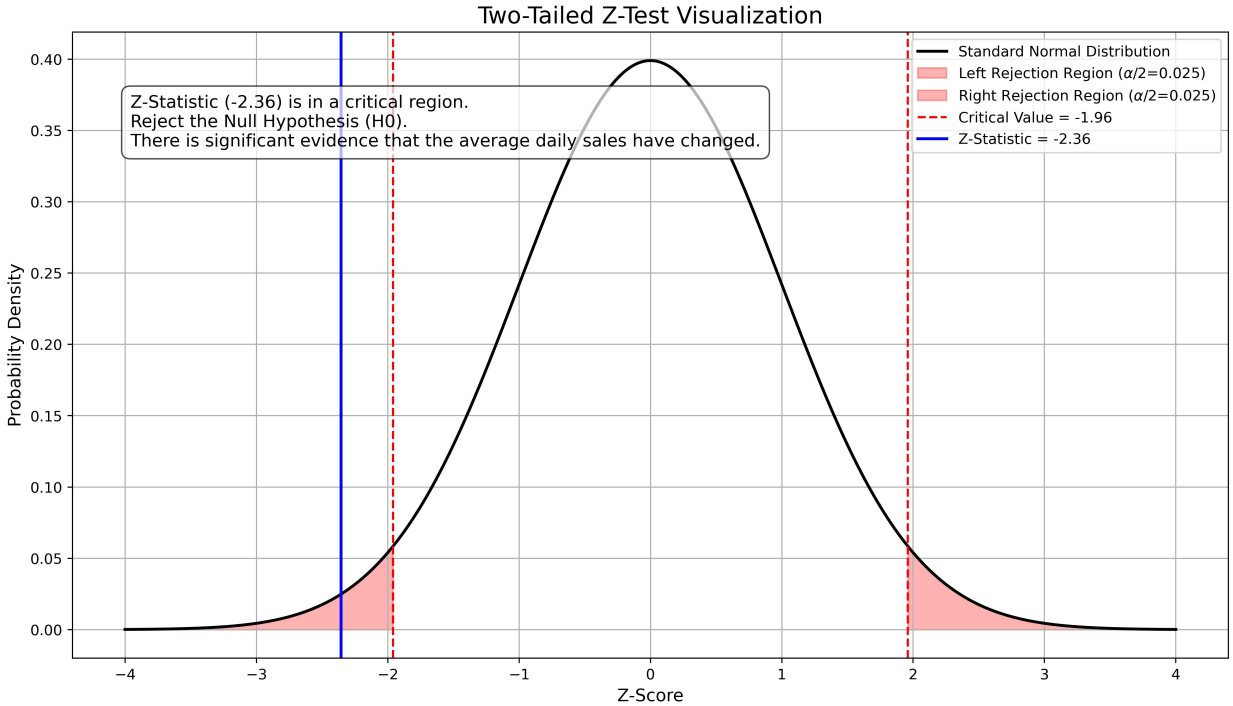

Figure: Here H_0 is that the daily sales of the shop are \$ 1200, while H_1 is that it is different from that

Example: Left-Tailed T-Test

Problem: A company claims its light bulbs last 1,000 hours. A sample of 15 bulbs shows an average life of 980 hours with a sample standard deviation of 75 hours. Is the company's claim false at a 1% significance level?

1. **Hypotheses:**
   - $H_0 : \mu = 1000$ (The claim is true)
   - $H_a : \mu < 1000$ (The claim is false)
2. **Significance Level:** $\alpha = 0.01$.
3. **Calculation:**
   - Standard Error: $SE = \frac{s}{\sqrt{n}} = \frac{75}{\sqrt{15}} \approx 19.36$
   - Test Statistic: $t = \frac{\bar{x}-\mu}{SE} = \frac{980-1000}{19.36} \approx -1.03$
   - Degrees of Freedom: $df = n - 1 = 14$
   - p-value (from t-distribution table or software): $P(t < -1.03) \approx 0.161$
4. **Decision:** Since p-value $(0.161) > \alpha$ (0.01), we **fail to reject** H_0.
5. **Conclusion:** There is not enough evidence to conclude that the company's claim is false.

Left-Tailed T-Test: Figure

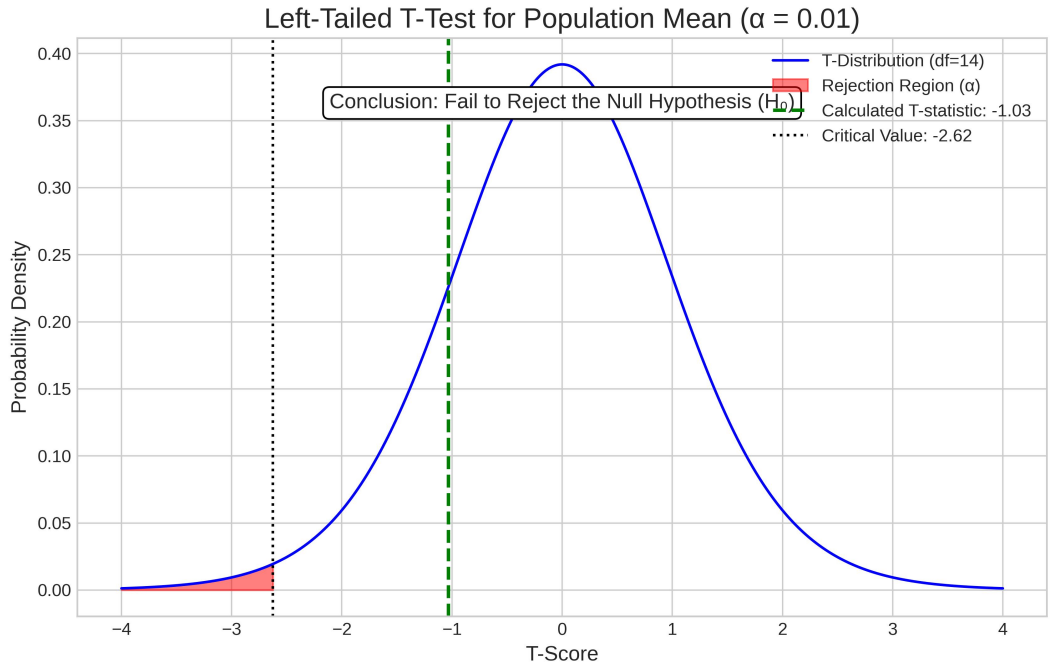

Figure: Here H_0 is that the daily sales of the shop are \$ 1200, while H_1 is that it is different from that

Example: A Right-Tailed T-test

Problem: A sports drink manufacturer claims that their product contains more than 50 grams of carbohydrates per serving. A nutritionist tests this claim by taking a sample of 15 servings. The sample has an average of 53 grams of carbohydrates with a standard deviation of 4 grams. The nutritionist wants to know if there is enough evidence to support the claim that the actual mean $\mu > 50$, at a 0.05 significance level.

Hypotheses:

- **Null Hypothesis** (H_0): $\mu \leq 50$ (The average carbohydrate content is less than or equal to 50 grams)
- **Alternative Hypothesis** (H_a): $\mu > 50$ (The average carbohydrate content is greater than 50 grams).

Right-Tailed T-test: solution

Here, $\mu = 50$, $\bar{x} = 53$, $s = 4$, $n = 15$, therefore

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{53 - 50}{\frac{4}{\sqrt{15}}} = 2.9047.$$

Now, $p(t = 2.9047) = 0.0058$, which implies $p < 0.05$, therefore, we reject the null hypothesis. Alternatively, $t_{crit} = t(0.05) = 1.7613 < t = 2.9047$ which also implies that H_0 is to be rejected and in all likelihood the true mean is more that 50. This is explained in the following graph.

Right-tailed T-test: Plot

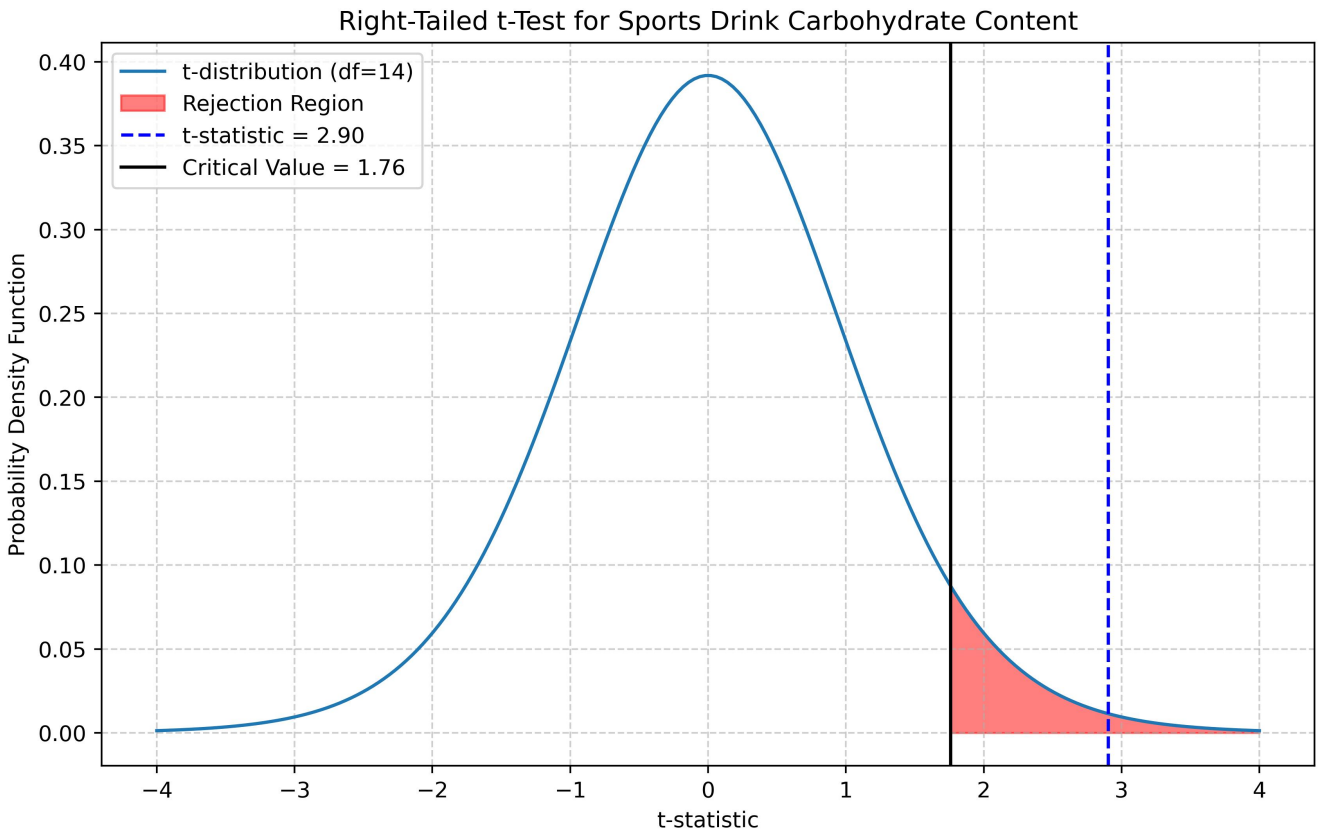

Figure: Here H_0 is that the mean carbohydrate content in a drink is less than 50 gm, while H_1 is that it is greater than that value. The p-value here is smaller than 0.05 (or $t_{crit} < t$), therefore, H_0 must be rejected.

Example: A two-tailed t-test

Problem Statement: A factory's manufacturing process is designed to produce cylindrical rods with a mean diameter of 2.50 cm. The quality control department suspects that the process has drifted and is no longer producing rods with this exact mean diameter. They take a random sample of 30 rods and measure their diameters. The sample mean is found to be 2.53 cm with a standard deviation of 0.08 cm.

Solution: We will perform a two-tailed t-test to determine if the sample mean is significantly different from the claimed population mean of 2.50 cm at a 5% significance level. The plot will show the t-distribution, the two rejection regions, and the calculated t-statistic.

The Two-tailed t-test...

Clearly

$$H_0 : \mu = 2.50 \qquad H_1 : \mu \neq 2.50$$

Here $\mu = 2.50$, $s = 0.08$, $n = 30$, $\bar{x} = 2.53$, $\alpha = 0.05$, therefore, the test statistic and the t value will be

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{2.53 - 2.50}{0.08/\sqrt{30}} = 2.054$$

Significance level: $\alpha = 0.05$, but $t_{crit} = t(\alpha/2) = \pm 2.0452$

- Critical values: ± 1.96 for t, therefore we reject H_0 because$|t| = 2.054 > t_{crit} = 2.045$
- Therefore, H_0 is to be rejected and indeed the diameters have changed

Plot of the two-tailed t test

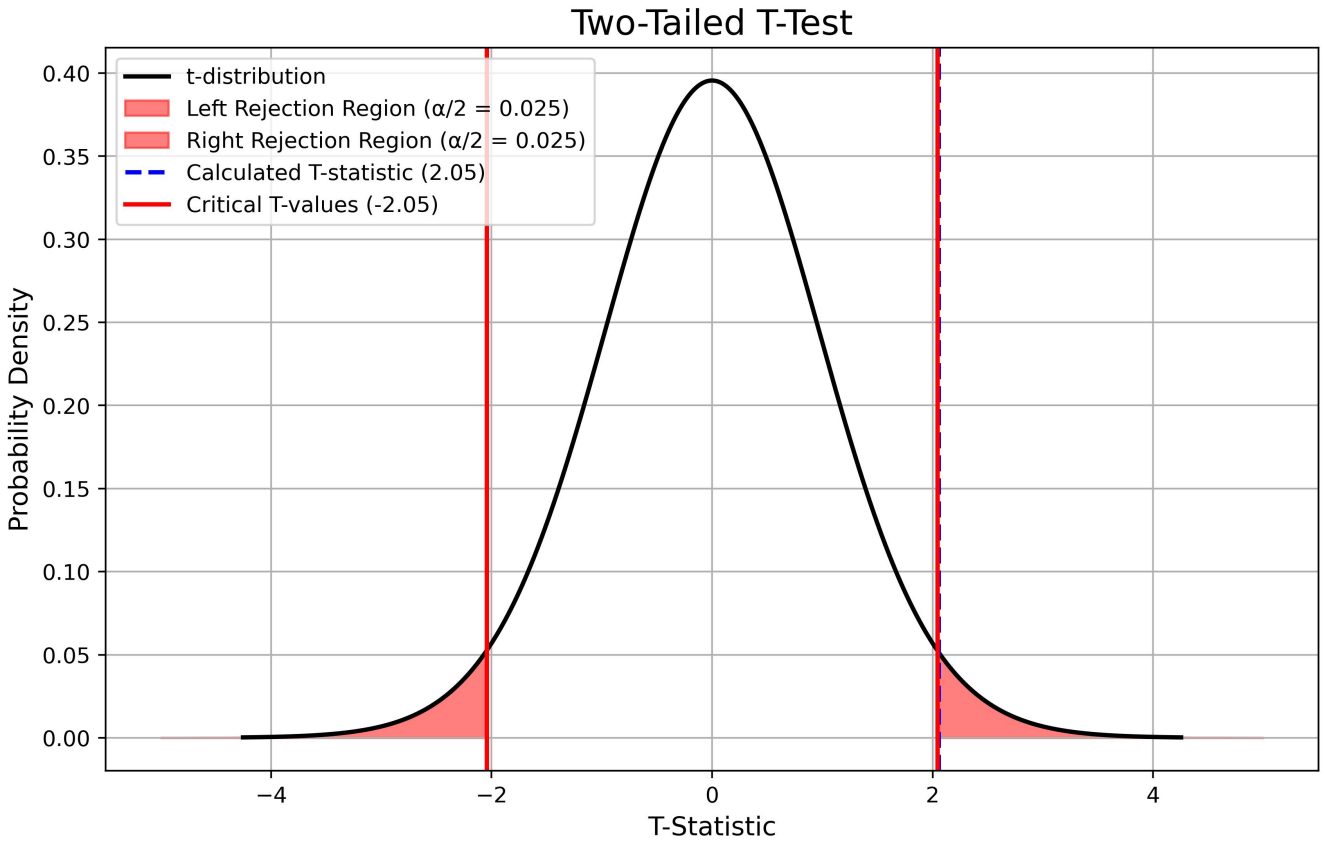

Figure: Here H_0 is that the diameters of the rods are 2.50 cm, while H_1 is that it is different from that

One-tailed vs. Two-tailed Tests

- Whether or not the average height of male students is $\mu = 175$ cm
- The alternative hypothesis determines if the test is one-tailed or two-tailed.
- **Two-tailed test ($H_a : \mu \neq 175$):** We are looking for a difference in either direction (greater than or less than). The p-value is calculated from both tails of the distribution.
- **One-tailed test ($H_a : \mu > 175$ or $H_a : \mu < 175$):** We are looking for a difference in a specific direction. The p-value is calculated from only one tail.

Example: Two-Sample T-Test

Problem: We want to compare the average test scores of two different classes. Class A (20 students) has a mean score of 85 and a standard deviation of 5. Class B (25 students) has a mean score of 88 and a standard deviation of 6. Is there a significant difference between the classes at $\alpha = 0.05$?

1. **Hypotheses:**
   - $H_0 : \mu_A = \mu_B$ (No difference in mean scores)
   - $H_a : \mu_A \neq \mu_B$ (Significant difference in mean scores)
2. **Significance Level:** $\alpha = 0.05$ (two-tailed).
3. **Calculation (assuming equal variances):**
   - Pooled Variance: $s_p^2 = \frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2} = \frac{19(5^2)+24(6^2)}{20+25-2} \approx 31.33$
   - Test Statistic: $t = \frac{(\bar{x}_A-\bar{x}_B)-(\mu_A-\mu_B)}{\sqrt{s_p^2(\frac{1}{n_A}+\frac{1}{n_B})}} = \frac{85-88}{\sqrt{31.33(\frac{1}{20}+\frac{1}{25})}} \approx -1.74$
   - Degrees of Freedom: $df = n_A + n_B - 2 = 43$.
   - p-value: $P(|t| > 1.74) \approx 0.088$
4. **Decision:** Since p-value $(0.088) > \alpha$ (0.05), we **fail to reject** H_0.
5. **Conclusion:** There is no statistically significant difference in the mean test scores of the two classes.

What is a Chi-Square Test?

- The Chi-Square (χ^2) test is a non-parametric statistical test.
- It's used to determine if there is a significant association between two categorical variables.
- It compares observed frequencies to expected frequencies under the assumption of the null hypothesis.

Key Concepts

- **Categorical Data:** Data that can be divided into distinct categories (e.g., gender, color, yes/no responses).
- **Observed Frequencies (O):** The actual counts from your sample data.
- **Expected Frequencies (E):** The counts you would expect to see if there were no relationship between the variables.
- **Chi-Square Goodness-of-Fit Test:**

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where O_i = Observed frequencies, E_i = Expected frequencies.

- **Chi-Square Test of Independence:**

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad E_{ij} = \frac{(\text{Row Total})_i (\text{Column Total})_j}{\text{Grand Total}},$$

above E_{ij} denotes the expected frequencies.

Steps for a Chi-Square Test

1. **State the Hypotheses:**
   - H_0: The variables are independent (no association).
   - H_a: The variables are dependent (a significant association exists).
2. **Calculate Expected Frequencies:** Based on the null hypothesis.
3. **Calculate the χ^2 statistic:** Using the formula.
4. **Determine the Degrees of Freedom (df):**
   - $df = (\text{rows} - 1) \times (\text{columns} - 1)$ for a test of independence.
5. **Find the p-value:** Using the χ^2 value and df.
6. **Draw a Conclusion:** Compare the p-value to the significance level (α). If $p < \alpha$, reject H_0.

Example: Goodness of Fit for the coin-toss problem

Problem: A coin is tossed 100 times, resulting in 65 heads and 35 tails. We want to test if the coin is fair at a 0.05 significance level.
Solution: We have

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$
$$= \frac{(65-50)^2}{50} + \frac{(35-50)^2}{50} = 9$$

This corresponds to a p-value $0.0027 < 0.05$, the significance level. Therefore, H_0 is rejected and we conclude that the coin is not fair.

Figure for the coin toss problem

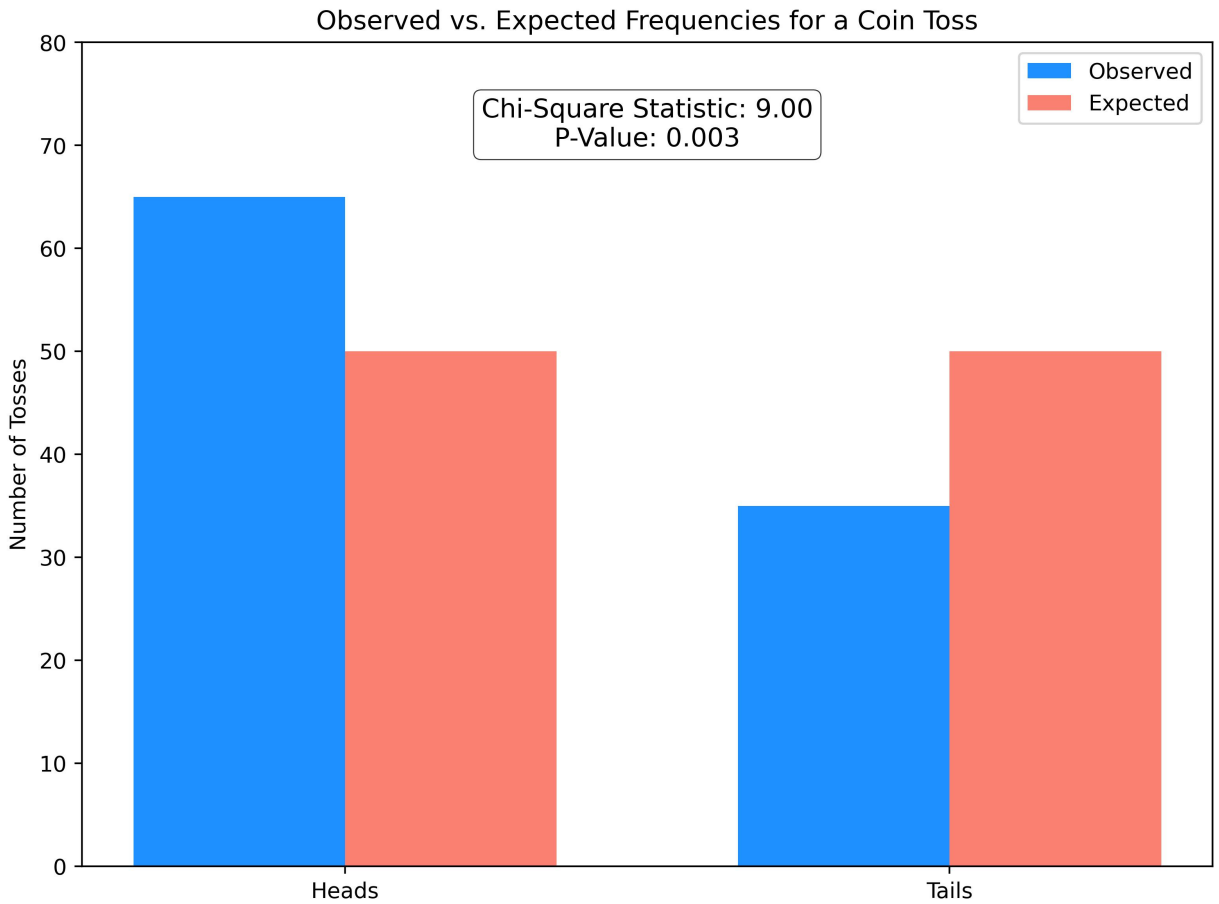

Figure: Here H_0 is that the coin is fair, while H_1 is that it is not fair. Our calculation reveals that H_0 is to be rejected.

Example 1: Goodness-of-Fit

A die is rolled 60 times. The outcomes are:

Face	1	2	3	4	5	6
Observed (O_i)	8	9	19	5	8	11
Expected (E_i)	10	10	10	10	10	10

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(8-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(19-10)^2}{10} + \cdots + \frac{(11-10)^2}{10}$$

Example 1: Solution

$$\chi^2 = \frac{4}{10} + \frac{1}{10} + \frac{81}{10} + \frac{25}{10} + \frac{4}{10} + \frac{1}{10}$$

$$\chi^2 = 11.6$$

- Degrees of freedom $= 6 - 1 = 5$
- Critical value at $\alpha = 0.05$ is 11.07
- Since $11.6 > 11.07$, we **reject** H_0.
- Conclusion: The die is not fair.

Example 2: Test of Independence

A survey was conducted on 100 people regarding gender and preference for a new product.

	Like	Dislike	Total
Male	20	30	50
Female	30	20	50
Total	50	50	100

$$E_{ij} = \frac{(i-\text{th Row Total})(j-\text{th Column Total})}{\text{Grand Total}}$$

Example 2: Expected Frequencies

$$E_{11} = \frac{50 \times 50}{100} = 25, \quad E_{12} = 25$$

$$E_{21} = 25, \quad E_{22} = 25$$

	Like	Dislike
Male	25	25
Female	25	25

Example 2: Chi-Square Calculation

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(20-25)^2}{25} + \frac{(30-25)^2}{25} + \frac{(30-25)^2}{25} + \frac{(20-25)^2}{25}$$

$$= \frac{25}{25} + \frac{25}{25} + \frac{25}{25} + \frac{25}{25} = 4$$

- Degrees of freedom $= (2-1)(2-1) = 1$
- Critical value at $\alpha = 0.05$ is 3.84
- Since $4 > 3.84$, reject H_0.
- Conclusion: Gender and preference are dependent.

Example 3: Chi-Squared Test

Problem: We want to know if there is a relationship between a person's gender and their preference for coffee or tea.

	Coffee	**Tea**	**Total**
Male	45	25	70
Female	35	55	90
Total	80	80	160

- **Hypotheses:**
  - H_0: Gender and beverage preference are independent.
  - H_a: Gender and beverage preference are not independent.

Chi-squared test...

- **Calculation:**
  - Expected count for Male/Coffee: $E = \frac{70 \times 80}{160} = 35$
    - Chi-Square Statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$
    - $\chi^2 = \frac{(45-35)^2}{35} + \frac{(25-35)^2}{35} + \frac{(35-45)^2}{45} + \frac{(55-45)^2}{45} \approx 2.86 + 2.86 + 2.22 + 2.22 \approx 10.16$
    - Degrees of Freedom: $df = (rows - 1)(cols - 1) = (2-1)(2-1) = 1$
    - p-value for $\chi^2 = 10.16$ and $df = 1$ is very small, $p \approx 0.0014$.
  - **Decision:** Since p-value $(0.0014) < \alpha$ (0.05), we **reject** H_0.
  - **Conclusion:** There is a significant relationship between gender and beverage preference.

Summary of Examples

- We saw how to apply the five-step process to different types of data.
- **Z-Test:** Large sample, known population standard deviation.
- **T-Test:** Small sample, unknown population standard deviation.
- **Chi-Square Test:** Categorical data.
- The key is to correctly identify the test and interpret the p-value relative to the significance level.

Summary

- Hypothesis testing is a formal procedure for making a decision about a population based on sample data.
- It involves stating a null and alternative hypothesis and using a p-value to decide whether to reject the null hypothesis.
- Understanding Type I and Type II errors is crucial for interpreting the results correctly.
- **Final thought:** Hypothesis testing helps us move from speculation to evidence-based conclusions.