Hypothesis Testing

Pratyush Kant

September 27, 2024

1 Introduction

Given a population (or distribution) with a parameter of interest, θ, the goal is to decide between two complementary statistical hypotheses concerning θ. The choice between these hypotheses is to be based on a sample data taken from the population.

1.1 Key Elements

1. **Parameter of Interest (θ)**: This could be the mean, variance, standard deviation, proportion, or any other relevant parameter of the population.

2. **Statistical Hypotheses**: Two complementary statements concerning the parameter of interest, θ.

3. **Sample Data**: Data taken from the population of interest, which will be used to make the decision between the hypotheses.

The ideal goal is to be able to choose the hypothesis that is true in reality based on the sample data. For instance, a quality engineer wants to determine whether the production process they are monitoring is still producing products with the required mean (process in-control), or if the mean response value is now different (process out-of-control).

The Hypotheses are:

1. **Null Hypothesis (H_0)**: This is the hypothesis that is assumed to be true unless there is sufficient evidence to the contrary. It is denoted by H_0. In this case, the null hypothesis is that the process is in-control, $\mu = \mu_0$.

2. **Alternative Hypothesis (H_1)**: This is the hypothesis that is accepted if the data provides sufficient evidence to reject the null hypothesis. It is denoted by H_1. The alternative hypothesis is that the process is out-of-control, $\mu \neq \mu_0$.

Where μ represents the mean response value of the products and μ_0 is the hypothesized mean response value.

We present some more examples of null and alternative hypotheses below:

- An engineer would like to decide which of two computer chip manufacturers (say, X and Y) is more reliable in producing computer chips. If we denote by p_1 the proportion of defective chips for X, and p_2 the proportion of defective chips for Y the:
  - Null Hypothesis: $H_0 : p_1 \leq p_2$, X is more reliable than Y.
  - Alternative Hypothesis: $H_1 : p_1 > p_2$, Y is more reliable than X.

1.2 Steps in Hypothesis Testing

1. **Formulate the Hypotheses**: Define the null and alternative hypotheses. H_0, is usually the hypothesis that corresponds to the "status quo", "the standard", "the desired level/amount", or it represents the statement of "no difference". The alternative hypothesis, H_1, on the other hand,

	Decision	**Decision**
Reality	**Accept H_0**	**Reject H_0**
H_0 is True	Correct Decision	Type I Error
H_0 is False	Type II Error	Correct Decision

Table 1: Types of Errors

is the complement of H_0, and is typically the statement that the researcher would like to prove or verify.[1]

These hypotheses are usually set-up in such a way that deciding in favor of H_1 when in fact H_0 is the true (called a Type I error) statement is a very serious mistake. The probability of making a Type I error is denoted by α.[2]

2. **Collect the Data**: Take a sample from the population. The researcher must decide on the sampling method, whether it is a simple random sample or a stratified sample. For this example, it is assumed that a simple random sample of size n will be obtained, where $n > 30$. The data will be represented as $X_1, X_2, \ldots, X_n$. It is also assumed that the population standard deviation σ is known.

3. **Decision Rule (Test Statistic)**: The decision rule is a criterion that is used to decide whether to reject the null hypothesis. The decision rule is based on a test statistic, which is a function of the sample data. The test statistic is used to determine the probability of observing the sample data, given that the null hypothesis is true. The test statistic is compared to a critical value, which is a value that is used to determine whether to reject the null hypothesis. The critical value is chosen based on the desired level of significance, α.

   A reasonable test statistic for testing the mean of a population when the population standard deviation is known is the z-statistic, which is defined as:

   $$z := \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

   where

   $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

   is the sample mean. The larger the value of $\mid z \mid$, the more evidence there is against the null hypothesis. The **critical value** denoted by C is the value that the test statistic must exceed in order to reject the null hypothesis.

2 Types of Errors

Error of Type I is considered to be a more serious type of error. Therefore, we try to minimize the probability of committing the Type I error. In trying to minimize, however, the probability of a Type I error, we encounter an obstacle in that the probabilities of the Type I and Type II errors are inversely related. Thus, if we try to make the probability of a Type I error very, very small, then it will make the probability of a Type II error quite large. As a compromise we therefore specify a **maximum tolerable Type I error probability**, called the **significance level**, and denoted by α, and choose the critical value C such that the probability of a Type I error is (at most) equal to α. This α is conventionally set to 0.10, 0.05, or 0.01.

$$\begin{aligned}\mathbb{P}(\text{Type I Error}) &= \mathbb{P}(\text{Reject } H_0 \mid H_0 \text{ is True})\\ &= \mathbb{P}(\mid z \mid > C \mid H_0 \text{ is True})\end{aligned}$$

[1] **A word of caution**: It is not proper for a researcher to set up the hypotheses after seeing the sample data; however, a data maybe used to generate a hypotheses, but to test these generated hypotheses you should gather a new set of sample data!

[2] The null hypothesis is usually the hypothesis that is assumed to be true unless there is sufficient evidence to the contrary.

Under the assumption that the null hypothesis is true, the test statistic z is normally distributed with mean 0 and standard deviation 1. Therefore, the probability of a Type I error is given by:

$$\mathbb{P}(\text{Type I Error}) = \mathbb{P}(\mid z \mid > C) = 2\mathbb{P}(z > C)$$

For $C = z_{\alpha/2}$, the critical value is such that $\mathbb{P}(z > z_{\alpha/2}) = \alpha/2$. Therefore, the probability of a Type I error is α. Hence the resulting decision rule is:

$$\text{Reject } H_0 \text{ if } \mid z \mid > z_{\alpha/2} \text{ for a significance level of } \alpha$$[3]

If $\mid z \mid < z_{\alpha/2}$, then we fail to reject the null hypothesis H_0.

When H_0 is rejected, then either that a correct decision has been made, or an error of Type I has been committed. But since we have controlled the probability of committing a Type I error (set to α), then we can conclude in this case that H_0 is not true, and hence that H_1 is correct.

3 The p-Value Approach

The p-value is defined as $\mathbb{P}(\mid z \mid > \mid z_\text{o} \mid)$, where z_o is the observed value of the test statistic. The p-value is the probability of observing a test statistic as extreme as the one computed from the sample data, assuming that the null hypothesis is true. The p-value is used to determine the strength of the evidence against the null hypothesis. The smaller the p-value, the stronger the evidence against the null hypothesis.[4]

The p-value is compared to the significance level, α, to determine whether to reject the null hypothesis. If the p-value is less than or equal to α, then the null hypothesis is rejected. If the p-value is greater than α, then the null hypothesis is not rejected.

1. If the p-value exceeds 0.10:
   - H_0 is not rejected
   - The result is considered not significant
2. If the p-value is between 0.10 and 0.05:
   - The result is considered almost significant or tending towards significance
3. If the p-value is between 0.05 and 0.01:
   - H_0 is rejected
   - The result is considered significant
4. If the p-value is less than 0.01:
   - H_0 is rejected
   - The result is considered highly significant
5. In general:
   - If the p-value is smaller than the chosen significance level (α), H_0 is rejected
   - If the p-value is larger than the chosen α, H_0 is not rejected
   - The smaller the p-value, the stronger the evidence against the null hypothesis

[3] The assumption is that σ is known and $n \geq 30$.

[4] As $\mid z_o \mid$ increases the p-value decreases, indicating stronger evidence against the null hypothesis.

4 Z-test

A z-test is used when we know the population standard deviation and the sample size is large (typically $n \geq 30$), or when we're working with proportions.

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

Where:

- $\bar{X}$ is the sample mean
- μ_0 is the hypothesized population mean
- σ is the known population standard deviation
- n is the sample size

Z-tests assume that the population is normally distributed and use the standard normal distribution. The z-score is compared to the critical value to determine whether to reject the null hypothesis. It is important to note that z-tests are more appropriate for large sample sizes and when the population standard deviation is known.

4.1 Example

Suppose we want to test if the mean height of adult males in a city is 170 cm. We know the population standard deviation is 10 cm. We take a sample of 100 men and find their mean height is 172 cm.

$$H_0 : \mu = 170 \text{ cm}$$
$$H_1 : \mu \neq 170 \text{ cm}$$

$$z = \frac{172 - 170}{10/\sqrt{100}} = 2$$

This z-score of 2 corresponds to a p-value of about 0.046 for a two-tailed test, which would be significant at $\alpha = 0.05$. Therefore, we reject the null hypothesis.

5 T-test

Overview A t-test is used when we don't know the population standard deviation, especially with smaller sample sizes. It's more commonly used in practice because we rarely know the population standard deviation.

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

Where:

- $\bar{X}$ is the sample mean
- μ_0 is the hypothesized population mean
- s is the sample standard deviation
- n is the sample size

T-tests use the t-distribution, which is similar to the normal distribution but accounts for the smaller sample size. The t-score is compared to the critical value to determine whether to reject the null hypothesis. It is important to note that t-tests are more appropriate for smaller sample sizes and when the population standard deviation is unknown.

There are three main types of t-tests:

1. One-sample t-test: Compares a sample mean to a known population mean
2. Independent samples t-test: Compares means of two independent groups
3. Paired samples t-test: Compares means of two related groups

5.1 Example (One-sample t-test)

Let's use the same height example, but now we don't know the population standard deviation. Our sample of 25 men has a mean height of 172 cm and a sample standard deviation of 8 cm.

$$H_0 : \mu = 170 \text{ cm}$$
$$H_1 : \mu \neq 170 \text{ cm}$$

$$t = \frac{172 - 170}{8/\sqrt{25}} = 1.25$$

With 24 degrees of freedom, this t-value corresponds to a p-value of about 0.223 for a two-tailed test, which would not be significant at $\alpha = 0.05$. Therefore, we fail to reject the null hypothesis.

5.2 Degrees of Freedom (df)

For a one-sample t-test:

$$df = n - 1$$

For an independent two-sample t-test:

$$df = n_1 + n_2 - 2$$

5.3 Standard Error (SE)

For a one-sample t-test:

$$SE = \frac{s}{\sqrt{n}}$$

For an independent two-sample t-test:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

5.4 Example: Independent Two-Sample T-Test

Is there a significant difference in test scores between two teaching methods?

- $H_0 : \mu_1 = \mu_2$ (no difference between means)
- $H_1 : \mu_1 \neq \mu_2$ (there is a difference between means)
- Method A: $n_1 = 30$, $\bar{x}_1 = 75$, $s_1 = 8$
- Method B: $n_2 = 30$, $\bar{x}_2 = 70$, $s_2 = 7$

Therefore, we have:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
$$t = \frac{75 - 70}{\sqrt{\frac{8^2}{30} + \frac{7^2}{30}}} \approx 2.582$$

The degrees of freedom are $df = n_1 + n_2 - 2 = 30 + 30 - 2 = 58$. For $\alpha = 0.05$ (two-tailed) and $df = 58$, $t_{crit} \approx 2.002$. Since, $|2.582| > 2.002$, so we reject the null hypothesis

6 Chi-Square Test

The chi-square test is used to determine whether there is a significant association between two categorical variables. It is a non-parametric test, which means it doesn't assume that the data is normally distributed. The chi-square test is used when the data is in the form of frequencies or counts for different categories.

6.1 One-Way Tables

The chi-squared distribution is used to test whether observed data differ significantly from theoretical expectations.

Suppose a die is rolled 36 times with the following frequencies:

Outcome	Frequency
1	8
2	5
3	9
4	2
5	7
6	5

$$\begin{aligned} H_0 &: \text{ Die is fair} \\ H_1 &: \text{ Die is not fair} \end{aligned}$$

The expected frequency for each value is $36 \times \frac{1}{6} = 6$. Compute

$$\sum \frac{(E-O)^2}{E}$$

where E is the expected frequency and O is the observed frequency. This sampling distribution of is approximately distributed as Chi Square with $k - m - 1$ degrees of freedom, where k is the number of categories and m eing the number of parameters being estimated from the data.

In the given example, $df = 6 - 0 - 1 = 5$ and $\chi^2(5) = 5.33$. Comparing it with the critical value C if $\chi^2(5) > C$ then reject H_0 else failed to reject H_0.

6.2 Contingency Tables

Contingency tables, also known as a cross-tabulation or crosstab, displays the frequency distribution of variables in a matrix format are used to examine the relationship between two categorical variables. The chi-square test of independence is used to determine whether there is a significant association between the two variables.

Let's consider a simple example studying the relationship between gender and preference for tea or coffee.

Table 2: Beverage Preference by Gender

	Beverage Preference		
Gender	Tea	Coffee	Total
Male	30	70	100
Female	50	50	100
Total	80	120	200

This contingency table shows:

- The sample consists of 200 individuals.
- There are equal numbers of males and females (100 each).
- 30 males prefer tea, while 70 prefer coffee.
- 50 females prefer tea, and 50 prefer coffee.
- In total, 80 individuals prefer tea, and 120 prefer coffee.

We can use this table to calculate various statistics:
Expected frequencies:

$$\begin{aligned}E(\text{Male, Tea}) &= \frac{100 \times 80}{200} = 40\\E(\text{Male, Coffee}) &= \frac{100 \times 120}{200} = 60\\E(\text{Female, Tea}) &= \frac{100 \times 80}{200} = 40\\E(\text{Female, Coffee}) &= \frac{100 \times 120}{200} = 60\end{aligned}$$

Chi-square calculation:

$$\begin{aligned}\chi^2 &= \frac{(30-40)^2}{40} + \frac{(70-60)^2}{60} + \frac{(50-40)^2}{40} + \frac{(50-60)^2}{60}\\&= 2.5 + 1.67 + 2.5 + 1.67\\&= 8.34\end{aligned}$$

With 1 degree of freedom ($(2-1)(2-1) = 1$), this χ^2 value is statistically significant at $p < 0.01$, indicating a significant association between gender and beverage preference. Notice that the degree of freedom is calculated as $(r-1)(c-1)$ where r is the number of rows and c is the number of columns.

It is worth noting that in the Chi Square test each subject contributes data to only one cell. Therefore, the sum of all cell frequencies in the table must be the same as the number of subjects in the experiment.

7 ANOVA (Analysis of Variance)

ANOVA is a statistical test used to analyze the differences among group means in a sample. It is used to compare the means of three or more groups to determine if there is a statistically significant difference between them. ANOVA tests the null hypothesis that the means of two or more groups are equal.

One-way ANOVA is used when there is one independent variable with two or more levels (groups) and one dependent variable. The independent variable is categorical, while the dependent variable is continuous.

Two-way ANOVA is used when there are two independent variables, and the interaction between these two variables is of interest.

Post-Hoc Testing:

- Pairwise Comparisons: After a significant ANOVA result, conduct multiple two-sample t-tests to compare each pair of groups. However, this approach increases the risk of Type I errors (false positives) due to multiple comparisons.

- Tukey's Honestly Significant Difference (HSD): To address the issue of multiple comparisons, Tukey's HSD test is used. This test controls the family-wise error rate and helps identify which specific pairs of groups differ significantly.

```python
# Importing library
from scipy.stats import f_oneway

# Weight of male infants fed with different types of infant formula
Type1 = [6.9, 6.8, 6.4, 7.1, 7.2]
Type2 = [6.3, 6.2, 6.4, 6.3, 6.1]
Type3 = [7.2, 6.5, 6.6, 7.0, 6.9]
Type4 = [7.1, 6.8, 7.1, 7.2, 6.7]

# Conduct the one-way ANOVA
f_oneway(Type1, Type2, Type3, Type4)
# Output: F_onewayResult(statistic=8.822222222222223, pvalue=0.0011050920264836543)
```

```python
# Importing library
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# Weight of male infants fed with different types of infant formula
```

```
Type1 = [6.9, 6.8, 6.4, 7.1, 7.2]
Type2 = [6.3, 6.2, 6.4, 6.3, 6.1]
Type3 = [7.2, 6.5, 6.6, 7.0, 6.9]
Type4 = [7.1, 6.8, 7.1, 7.2, 6.7]

# Combine data into a format suitable for Tukey’s HSD
data = Type1 + Type2 + Type3 + Type4
groups = [’Type1’]*len(Type1) + [’Type2’]*len(Type2) + [’Type3’]*len(Type3) + [’Type4’]*len(Type4)

# Conduct Tukey’s HSD
thsd = pairwise_tukeyhsd(endog=data, groups=groups)
print(thsd)
# Output: TukeyHSDResults
```

8 General Remarks

8.1 Formulating Hypotheses

- Null Hypothesis (H_0): Testable, measurable, and based on prior knowledge. States no effect or difference.

- Alternative Hypothesis (H_1 or H_a): Predicts an expected effect or difference.

- Types in ML/AI: Performance (accuracy, speed), user behavior (engagement, retention), business impact (conversion rates, revenue), ethical (fairness, bias reduction).

- Characteristics: Specific, clearly defined, measurable, testable, relevant and based on prior knowledge or theoretical framework.

8.2 Designing Experiments

- Experimental Units: Individual users, sessions, groups, clusters, or time periods.

- Controlling for Confounding Variables: Identify potential confounders, use control groups, and apply blocking, stratification, and crossover designs.

- Sample Size: Determine using power analysis, effect size, significance level (α), power ($1 - \beta$), and adjust for dropout or noise.

- Randomization: Apply simple, stratified, cluster, or adaptive randomization techniques.

- Test Duration: Consider seasonality, allow novelty effects to wear off, and balance speed with reliability.

- Minimizing Interference Between Groups: Prevent contamination, isolate effects in networked settings, and adhere to SUTVA.

- Ethical Considerations: Ensure informed consent, protect user privacy, ensure fairness in treatment allocation, and minimize negative impacts.

8.3 SUTVA (Stable Unit Treatment Value Assumption)

- Assumption: Treatment to one unit does not affect others’ outcomes.

- Importance: Essential for valid causal inferences in A/B testing and ensuring that the effect measured is solely due to the treatment.

- Challenges in ML/AI: Network effects, shared resources, and adaptive algorithms.

- Strategies to Maintain SUTVA: Isolate test groups, use non-overlapping time periods, and design clustered experiments.

- Violations: Can bias treatment effect estimates and may necessitate complex experimental designs or analysis.

8.4 Sample Size and Statistical Power

Let β is the Type II error rate. Then,

$$\text{Power} = 1 - \beta = \mathbb{P}(\text{Reject } H_0 \mid H_1 \text{ is true})$$

Sample Size Calculation

Basic formula for a two-sample t-test:

$$n = \frac{2\sigma^2(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{d^2}$$

Where:

- n = sample size per group
- σ = population standard deviation
- d = effect size
- $z_{1-\frac{\alpha}{2}}$ = critical value for significance level
- $z_{1-\beta}$ = critical value for power

8.5 Importance of Statistical Tests

Statistical tests are essential for identifying relationships between variables in a dataset. For instance, when examining categorical variables (e.g., gender, education level) and numerical variables (e.g., salary), tests can determine:

- Whether there is a significant difference in mean salary between genders (using a 2-sample *t*-test).
- Whether mean salary differs across education levels (using a 1-way ANOVA).
- Whether mean salary varies by both gender and education level (using a 2-way ANOVA).

Statistical tests also evaluate effectiveness in various contexts, such as assessing treatment efficacy, evaluating educational course impacts, or measuring improvements from new ML models. Control trials and A/B testing are employed to generate data for such hypothesis testing.

8.6 Hierarchy of Tests

- Binary Categorical Variable Across Levels of Another Binary Categorical Variable: Use a 2-sample z-test (not a t-test) to compare the rate of a binary outcome (e.g., heart attack occurrence) between two groups (e.g., aspirin vs. no aspirin).
- Categorical Variable Across Levels of a Binary Categorical Variable: Employ the Chi-Squared Test to assess the distribution of a categorical outcome (e.g., heart attack severity with multiple levels) between two groups (e.g., aspirin vs. no aspirin).
- Categorical Variable Across Levels of Another Categorical Variable: Use the Chi-Squared Test to evaluate the distribution of a categorical outcome (e.g., heart attack severity with multiple levels) across different categories of a second variable (e.g., different aspirin dosages).
- Numerical Variable Across Levels of a Binary Categorical Variable: Apply a 2-sample z-test or t-test to compare a numerical outcome (e.g., weight) between two groups (e.g., government vs. private primary schools).
- Numerical Variable Across Levels of One Categorical Variable: Use ANOVA or a 1-way ANOVA test to compare a numerical outcome (e.g., weight) across different categories of a single categorical variable (e.g., different school syllabi).

- Numerical Variable Across Levels of Two Categorical Variables: Apply a 2-way ANOVA to analyze a numerical outcome (e.g., weight) across combinations of two categorical variables (e.g., rural vs. urban and government vs. private schools), including main effects and interactions.
- Numerical Variable Across Multiple Levels of One or More Variables: Utilize Linear Regression Analysis to explore the dependence of a numerical outcome (e.g., weight) on multiple predictors (e.g., age, height, calories consumed).
- Categorical Variable Across Multiple Levels of One or More Variables: Use Logistic Regression Analysis to examine the impact of multiple predictors (e.g., age, height, diet type) on a categorical outcome (e.g., heart attack incidence).