[Company name]

Test Statistics

[Document subtitle]

Table of Contents

Al-amine Mouhamad

1-1-2024

Introduction to Test Statistics

What is Test Statistics?

Test statistics are numerical values derived from sample data during a hypothesis test. They are used to determine whether to reject the null hypothesis (H_0) in favor of the alternative hypothesis (H_a).

Why is Test Statistics Important?

- **In Data Science:** Validates models and interprets trends.
- **In AI Research:** Evaluates model performance.
- **In Bioinformatics:** Analyzes gene expression and biological data.

Types of Testing

1. **Hypothesis Testing:** Framework to decide between H_0 and H_a.
2. **Parametric vs. Non-Parametric Tests:**
   - Parametric tests assume data follows specific distributions (e.g., normal).
   - Non-parametric tests make no such assumptions.

Hypothesis Testing

Key Concepts

Null Hypothesis (H_0) and Alternative Hypothesis (H_a)

* **H_0:** No effect or difference exists.

* **H_a:** Effect or difference exists.

Significance Level (α)

The probability of rejecting H_0 when it is true. Common levels: 0.05, 0.01.

p-value

The probability of observing the test results under H_0. * If p-value $\leq \alpha$, reject H_0.

Steps in Hypothesis Testing

1. Define hypotheses H_0 and H_a.
2. Choose an appropriate test and check assumptions.
3. Calculate the test statistic.
4. Compare the test statistic or p-value against critical values.
5. Conclude based on α.

Types of Errors

- **Type I Error (α):** Rejecting H_0 when it is true.

- **Type II Error (β):** Failing to reject H_0 when H_a is true (the probability of failing to reject H_0 when H_a is true).

Interpreting p-values and Statistical Significance

- A p-value less than the chosen significance level (α) indicates strong evidence against H_0, suggesting that the observed difference is statistically significant.
- A p-value greater than α suggests that there is not enough evidence to reject H_0, and the observed difference may be due to chance.

Parametric Tests

t-tests

One-Sample t-test

Tests whether the sample mean differs from a known value.

Formula: $t = (\bar{x} - \mu) / (s / \sqrt{n})$

Where:

- $\bar{x}$: Sample mean
- μ: Hypothesized population mean
- s: Sample standard deviation
- n: Sample size

Assumptions:

- Independence of observations
- Normality of the data or a large sample size ($n > 30$)

Python Example:

```
from scipy.stats import ttest_1samp
import numpy as np

data = [1.1, 2.2, 3.1, 4.3, 5.5]
t_stat, p_value = ttest_1samp(data, popmean=3.0)
print(f"T-statistic: {t_stat}, P-value: {p_value}")
```

R Example:

```
data <- c(1.1, 2.2, 3.1, 4.3, 5.5)
t.test(data, mu=3.0)
```

Independent Two-Sample t-test

Compares means of two independent groups.

Assumptions:

- Independence of observations
- Normality of the data or a large sample size (n > 30) for each group
- Equal variances between the two groups

Paired t-test

Compares means of paired data.

Non-Parametric Tests

Mann-Whitney U Test

Used to compare two independent samples.

Python Example:

```
from scipy.stats import mannwhitneyu

group1 = [1, 2, 3, 4, 5]
group2 = [6, 7, 8, 9, 10]
u_stat, p_value = mannwhitneyu(group1, group2)
print(f"U-statistic: {u_stat}, P-value: {p_value}")
```

R Example:

```
group1 <- c(1, 2, 3, 4, 5)
group2 <- c(6, 7, 8, 9, 10)
wilcox.test(group1, group2)
```

Chi-Square Tests

Test for Independence

Checks if two categorical variables are independent.

Python Example:

```
import numpy as np
from scipy.stats import chi2_contingency

data = np.array([[10, 20, 30], [6, 9, 17]])
chi2, p, dof, expected = chi2_contingency(data)
print(f"Chi2: {chi2}, P-value: {p}")
```

R Example:

```r
data <- matrix(c(10, 20, 30, 6, 9, 17), nrow=2)
chisq.test(data)
```

Correlation Tests

Pearson Correlation

Measures linear relationship between two variables.

Interpretation of the correlation coefficient (r):

- r = 1: Perfect positive linear relationship
- r = -1: Perfect negative linear relationship
- r = 0: No linear relationship

Python Example:

```python
from scipy.stats import pearsonr

x = [1, 2, 3, 4, 5]
y = [5, 6, 7, 8, 7]
r, p_value = pearsonr(x, y)
print(f"Correlation: {r}, P-value: {p_value}")
```

R Example:

```r
x <- c(1, 2, 3, 4, 5)
y <- c(5, 6, 7, 8, 7)
cor.test(x, y)
```

Projects and Assignments

1. **Gene Expression Analysis (Bioinformatics):** Use ANOVA to identify significant differences in gene expression across conditions.
2. **A/B Testing (Data Science):** Evaluate the effectiveness of a website feature using t-tests.
3. **Model Validation (AI Research):** Use resampling methods to validate a machine learning model's performance.

	Z-test	T-test	F-test	Chi-square
What does it test?	A single population mean	A single population mean	Equality of 2 population variances	A single population variance
Ha options	Ha: $\mu \neq \#$ Ha: $\mu > \#$ Ha: $\mu < \#$	Ha: $\mu \neq \#$ Ha: $\mu > \#$ Ha: $\mu < \#$	Ha: $\sigma_1^2 \neq \sigma_2^2$ Ha: $\sigma_1^2 > \sigma_2^2$ Ha: $\sigma_1^2 < \sigma_2^2$	Ha: $\sigma_1^2 \neq \#$ Ha: $\sigma_1^2 > \#$ Ha: $\sigma_1^2 < \#$
Other requirements	n.a	n.a	- Independent samples - Normal distrib.	Normal distribution
Critical value ◇	Z-table α left probabilities	T-table df = n-1 Upper tail probs.	F-table $df_1 = n_1 - 1$ $df_2 = n_2 - 1$	Chi-square table df = n-1
Test-statistic ★	$(\bar{x} - \mu) \div \left(\frac{\sigma}{\sqrt{n}}\right)$	$(\bar{x} - \mu) \div \left(\frac{s}{\sqrt{n}}\right)$	$\left(\frac{s_1^2}{s_2^2}\right)$ $\leftarrow$ larger variance	$(n-1)(s^2) \div (\sigma^2)$

	Diff. in means	Mean differences	Pearson correl.
What does it test?	Equality of 2 means (pop. variances assumed =)	Mean of the differences (paired comparisons)	Population correlation coefficient
Ha options	Ha: $\mu_1 - \mu_2 \neq 0$ Ha: $\mu_1 - \mu_2 > 0$ Ha: $\mu_1 - \mu_2 < 0$	Ha: $\mu_d \neq \#$ Ha: $\mu_d > \#$ Ha: $\mu_d < \#$	$\rho \neq 0$
Other requirements	- Independent samples - Normal distrib.	- DEPENDENT samples - Normal distrib.	- Normal distrib.
Critical value ◇	T-table $df = n_1 + n_2 - 2$	T-table df = n-1	T-table df = n - 2
Test-statistic ★	$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\left(\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}\right)^{1/2}}$	$(\bar{d} - \mu_d) \div \left(\frac{Sd}{\sqrt{n}}\right)$	$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$