Statistics Notes and Practice

Descriptive Statistics and Measures of Central Tendency

Descriptive statistics summarize and describe the main features of a dataset. Measures of central tendency (mean, median, and mode) represent the central point of the data.

Skewness and Kurtosis

- Skewness measures the asymmetry of the data distribution. For symmetrical data, skewness is 0. Positive skewness means a long right tail, and negative skewness means a long left tail.

- Kurtosis measures the 'tailedness' of the data. For symmetrical data, kurtosis is 3 (mesokurtic). Higher kurtosis (>3) indicates heavy tails (leptokurtic), and lower kurtosis (<3) indicates light tails (platykurtic).

Hypothesis Testing

Hypothesis testing determines whether there is enough evidence in a sample to infer that a condition holds for the entire population.

- Purpose: To test assumptions (hypotheses) using statistical methods.
- Types:
  - Z-Test
  - T-Test
  - Chi-Square Test
  - ANOVA

Errors in Hypothesis Testing:
- Type I Error: Rejecting a true null hypothesis (false positive).
- Type II Error: Failing to reject a false null hypothesis (false negative).
Used in regression for parameter significance and in classification for threshold optimization.

Methods and Python Code for Hypothesis Testing

1. Z-Test

Used When: Large sample size ($n > 30$), known population variance.

Steps:
1. State the null and alternative hypotheses.
2. Calculate the test statistic:
   $Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$
3. Compare the Z-value to the critical value or use the p-value.
4. Decide to accept or reject the null hypothesis.

Python Code:

```
from scipy.stats import norm
sample_mean = 50
population_mean = 45
std_dev = 10
sample_size = 36

z_score = (sample_mean - population_mean) / (std_dev / (sample_size ** 0.5))
critical_value = norm.ppf(1 - alpha / 2)  # Two-tailed critical value
p_value = 2 * (1 - norm.cdf(abs(z_score)))
```

2. t-Test

Used When: Small sample size ($n \leq 30$), unknown population variance.

Steps:
1. State the null and alternative hypotheses.
2. Calculate the test statistic:
   $t = (\bar{X} - \mu) / (S / \sqrt{n})$
3. Compare the t-value with the critical t-value or p-value.

Refer ENTRI in app videos for t test practice

p value:

The p-value tells us how likely it is to get the observed results if the null hypothesis (H_0) is true. It is compared to the significance level (α), which is a threshold set before testing to decide whether to reject H_0.
If the p-value is less than or equal to α, we reject H_0, indicating significant evidence for the alternative hypothesis (H_1). If the p-value is greater than α, we fail to reject H_0, meaning there's insufficient evidence against it.
The significance level α represents the maximum risk of making a wrong decision (Type I error). Common values for α are 0.05 (5%) or 0.01 (1%). While a smaller α (e.g., 0.01) reduces the chance of rejecting H_0 incorrectly. The p-value shows the confidence in the results up to the chosen α.

p-value in a Z-Test

```
from scipy.stats import norm
# Example: Z-test
z_score = 1.96  # Replace with your calculated z-score
p_value = 2 * (1 - norm.cdf(abs(z_score)))  # Two-tailed test
print("P-value for Z-test:", p_value)
```

p-value in a t-Test

```
from scipy.stats import t# Example: T-test
t_score = 2.1  # Replace with your calculated t-score
degrees_of_freedom = 10  # Replace with your sample's degrees of freedom
p_value = 2 * (1 - t.cdf(abs(t_score), df=degrees_of_freedom))  # Two-tailed test
print("P-value for T-test:", p_value)
```

*The **degrees of freedom (df)** for a sample are calculated using the formula:*
$df=n-1$

Where n is the sample size (the total number of observations in the sample).

This formula is used in most t-tests and reflects the number of independent data points available for estimating variability.

Z-Test vs. Z-Score

- Z-Test Equation:
  $Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$
  Used in hypothesis testing.

- Z-Score (Outlier Detection):
  $Z = (X - \mu) / \sigma$
  Measures how many standard deviations a data point is from the mean.

Practice Activity

1. Descriptive Statistics: Calculate the mean, median, and mode of the dataset: [12, 15, 14, 10, 8, 14, 10].
2. Skewness and Kurtosis: Interpret the skewness = 1.2 and kurtosis = 2.8 for a dataset.
3. Z-Test: Perform a Z-test for a sample mean of 25, population mean of 20, standard deviation of 5, and sample size of 30. The confidence interval is 95%.
4. T-Test: Conduct a t-test for the dataset [3.4, 3.7, 3.1, 3.6, 3.3] with a population mean of 3.5. The confidence interval is 95%.

Answers:

1. Mean: 12.14, Median: 12, Mode: 10 and 14.
2. Positively skewed; light tails (platykurtic).
3. Z = 5.48, P-value < 0.05 (reject null hypothesis).
4. t = -0.44, P-value > 0.05 (fail to reject null hypothesis).