**Assignment #3: Instructor Assignment –**

**Practical Statistics**

**COM 968-32: Statistics for Social Research II**

**(Fall 2024, Sub-term A)**

**Dr. Peter Abraham Airewele, DSL**

**Omega Graduate School**

**Professor**

**Dr. Sean Taladay, EdD**

**November 12, 2024**

**Assignment #3: Instructor Assignment - Practical Statistics**

Answer the following questions in an essay format, with 1-2 fully developed paragraphs for each question. Include citations/references from your Developmental Reading log.

1. What is normality? Why is it relevant to parametric (generalizable to the target population) versus nonparametric (applicable only to the sample) statistics in social science research?
2. What is a histogram? What is a box and whisker plot? How are they useful for understanding the normality of a given sample?
3. Describe the following statistical procedures:

\* Quasi-Experimental: t-Tests (Independent and Dependent) - Parametric

\* Quasi-Experimental: Mann Whitney U and Wilcoxon - Nonparametric

\* Correlational: Pearson’s r - Parametric

\* Correlational: Spearman’s Rank - Parametric

4) Navigate to OGS’s Practical Statistics for Social Research (PSSR) tool. Click on

“Example Datasets” is a feature in OGS's Practical Statistics for Social Research (PSSR) tool that provides pre-loaded datasets for practice and learning. Load the “Independent t-Test: Ethical Decision-Making.” Dataset. Scroll to "Step Three: Run Statistical Procedures" and click "t-Test."

Dataset. Scroll to “Step Three: Run Statistical Procedures” and click “t-Test”.

* 1. **Copy and paste** the output's contents into your assignment document. Please read it carefully and

expand on it based on your understanding. Answer the following questions:

\* What might be good problem and purpose statements for this dataset?

\* What might be good research questions related to the hypotheses generated by

the PSSR software?

\* What does the output tell you about comparing the two groups?

\* How was the p-value used to test the hypotheses?

**4.2 Repeat steps 10-12 for** the “Correlational: Life Satisfaction Index” dataset. Note

that this is a correlational design with two continuous variables. What do the

scatterplots tell you about the relationship between the two variables?

**4.3 Repeat steps 10-12 for** both datasets, but use the nonparametric

equivalents of the statistical tests (Mann \* \* \* Whitney U and Spearman’s Rank,

respectively).

How did the results change?

For each dataset, click on “Assumptions” under “Step Two: Run Descriptives

and Assumptions.”

**4.4** Based on the output of each dataset, should you use parametric or nonparametric

procedures? Are the datasets usually distributed?

Note that this is a judgment call on behalf of the researcher and not a black-and-

white decision.

1. Finally, navigate OGS’s Practical Statistics for Social Research (PSSR) tool. Click

on “Example Datasets” and load the “Example: Perfect Correlation” dataset. Scroll to “Step Three: Run Statistical Procedures” and click “Linear Regression”. What does the scatterplot graph show you about the relationship between the X and Y variables?

1. Summarize what you learned from conducting these statistical tests.

Include a title page, well-developed introduction and conclusion paragraphs, a references page, and an in-text APA formatted to support your responses.

**Assignment #3: Instructor Assignment - Practical Statistics**

**Introduction**

Statistical analysis is a cornerstone of social science research, providing the tools to uncover patterns and draw meaningful conclusions. A fundamental assumption in many statistical tests is the normality of the data distribution. This study delves into the concept of normality (Shatz, 2024; Kieu & Minh, 2024), its characteristics, and its implications for statistical analysis. Moreover, this study explores the concept and definition of Normality, its visual representation as a bell-shaped curve, and its significance in statistical inference. Other areas of study interests will include the Parametric vs. Nonparametric Tests to understand the distinction between these two types of statistical, analytical tests (Kronthaler & Zöllner, 2021; Kumar, 2024) and the role of the normality assumption in parametric tests. Histograms, box plots' visual inspection properties, and statistical tests like the Shapiro-Wilk test will be examined. The Roles of Statistical Software like SPSS, R-Programming, and Python (Kronthaler & Zöllner, 2021; Kumar, 2024) can be used to assess normality, perform appropriate statistical tests, and visualize data will be determined. Furthermore, there will be an investigation of the comparative analysis of parametric and nonparametric statistical tests commonly employed in quasi-experimental and correlational research. The researcher will examine the assumptions, strengths, and limitations of parametric statistical tests, such as t-tests, ANOVA, and linear regression, including the t-tests (independent and dependent samples), Mann-Whitney U tests, Wilcoxon signed-rank tests, Pearson's correlation coefficient, and Spearman's rank correlation coefficient. By understanding the nuances of these tests, one can make informed decisions about the most appropriate statistical analysis for their specific research questions.

**1)** What is normality? Why is it relevant to parametric (generalizable to the target population) versus nonparametric (applicable only to the sample) statistics in social science research?

Understanding normality in Statistics is a theoretical concept and a practical tool that will guide your research and analysis. It refers to data distribution following a specific pattern known as the normal distribution. This distribution is characterized by its bell-shaped curve, where most data points cluster around the mean, and the frequency of data points decreases symmetrically on either side.

Relevance to Parametric and Nonparametric Statistics:

Some parametric statistical tests, such as t-tests, ANOVA, and linear regression, assume the data is normally distributed. This assumption is crucial because it allows researchers to make inferences about the population parameters based on sample statistics. The robustness of these tests, especially when the sample size is large, provides confidence in the analysis's results.

On the other hand, nonparametric tests do not require the assumption of normality. They can be used when the data is not normally distributed or when the sample size is small. While nonparametric tests are more flexible, they are generally less powerful than parametric tests. This means that they may be less likely to detect significant differences or relationships, especially when the sample size is small. By understanding the concept of normality and its implications for statistical analysis (Yang, 2013; Yashaswi, 2024; Sileyew, 2019), social scientists can make informed decisions about the appropriate statistical techniques to use in their research.

**2)** What is a histogram? What is a box and whisker plot? How are they useful for Histograms?

Box and whisker plots are potent tools for visualizing and understanding data distributions, including their normality.   A histogram is a graphical representation of the distribution of numerical data. It divides the data into bins or intervals and shows the frequency of data points falling within each bin.   The shape of the histogram can provide clues (Ose, 2016; Privitera, 2024) about the distribution's normality. A normal distribution typically appears bell-shaped, with most data points clustered around the mean and tapering off symmetrically on both sides.

A box and whisker plots are a graphical representation of the five-number summary of a dataset: i) the minimum value, ii) the first quartile (Q1), iii) the median (Q2), iv) the third quartile (Q3), v) and the maximum value.

The box represents the interquartile range (IQR), which contains the middle 50% of the data.

The whiskers extend from the box to the minimum and maximum values, excluding outliers.

A box plot can help identify potential outliers and skewness in the data. A normally distributed dataset typically has a symmetric box plot with roughly equal-length whiskers. Both histograms and box plots can assess a dataset's normality.

In a histogram, the bell-shaped curve suggests normality, and the skewness (asymmetry) indicates non-normality.  The multiple peaks (multimodality) also suggest non-normality.

In the Box Plot: A symmetric box plot with roughly equal-length whiskers suggests normality.

Skewness: A skewed box plot with one whisker that is significantly longer than the other indicates non-normality. The Outliers can distort the distribution and affect normality.   While these visual methods can provide initial insights, statistical tests like the Shapiro-Wilk or Kolmogorov-Smirnov tests can formally assess normality. However, it is essential to remember that no real-world data is usually distributed. Often, slight deviations from normality are acceptable, especially for larger sample sizes.

**3)**  Describe the following statistical procedures:

A) Quasi-Experimental Designs and t-Tests:

A Parametric Approach: Quasi-experimental designs are research methods that aim to establish a cause-and-effect relationship between variables without complete control of an actual experiment. While they lack random assignments, they often provide valuable insights into real-world phenomena. The T-tests are statistical tests used to compare the means of two groups (Ravid, 2024; Plodder & Hamann, 2021). They are instrumental in quasi-experimental designs when the dependent variable is continuous. The types of t-tests for quasi-experimental designs include: i) Independent Samples: The t-test compares the means of two independent groups. b) It compares the performance of two groups (e.g., treatment vs. control) that were not randomly assigned. c) It analyzes the pre-test and post-test scores of two different groups. ii) The Dependent Samples t-Test (Paired Samples t-Test): a) Compares the means of two related groups. b) The Quasi-Experimental Application analyzes the same group's pre-test and post-test scores and participants' performance on two different tasks or conditions.

The Assumptions of the t-test ensure the validity of t-test results, and the following assumptions must be met: a) Each observation should be independent of the others. b) The dependent variable should be normally distributed in each group. c) The variances of the two groups should be equal.

In Conducting t-tests in Quasi-Experimental Designs, the data is collected on the dependent variable for the two groups being compared. It uses statistical software like SPSS, R-programming, and Python to perform the appropriate t-test.

In the Hypothesis Testing, the Null Hypothesis (H₀): There is no significant difference between the means of the two groups. In the Alternative Hypothesis (H₁): There is a significant difference between the means of the two groups.

The p-value: If the p-value is less than the significance level (e.g., 0.05), the null hypothesis is rejected and concludes that there is a significant difference between the groups. However, there are limitations of t-tests in Quasi-Experimental Designs (Hadfield et al., 2022; Hatcher, 2013). Even when they may be powerful tools, their application in quasi-experimental designs has limitations: i) There is a Lack of Random Assignment: This can lead to potential biases and confounding variables. ii) There are threats to Internal Validity (Kelter, 2024; Litt, 2024), and factors such as selection bias, maturation, history, and testing effects can influence the results. iii) The findings may not be generalizable to a broader population. To mitigate these limitations, researchers often employ rigorous data analysis techniques, careful interpretation of results, and sensitivity analysis to assess the robustness of their findings. Researchers can draw meaningful conclusions from their studies by understanding the principles and limitations of t-tests in quasi-experimental designs.

B) Quasi-Experimental: Mann Whitney U and Wilcoxon – Nonparametric:

Nonparametric tests are statistical tests that do not assume the data follows a specific distribution, like the normal distribution. This makes them versatile for analyzing data that does not meet the assumptions of parametric tests.

The Mann-Whitney U test is a nonparametric test (MacFarland et al., 2016) used to compare two independent groups. It assesses whether the two groups differ in terms of their central tendency. The two groups being compared are independent of each other. It does not assume the data is usually distributed. It compares the central tendency of the two groups, often interpreted as comparing medians. The null hypothesis states that the two groups are identical.

The Wilcoxon signed-rank test is a nonparametric test used to compare two related samples. It is often used to analyze paired data, such as before-and-after measurements on the same subjects. The two groups are related, such as paired data. It does not assume the data is normally distributed. It compares the central tendency of the two related groups. The null hypothesis states that the difference between the two related groups is zero.

Interpreting Results: Both tests provide a p-value, which is the probability of observing the data or more extreme data if the null hypothesis is true. If the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant difference between the two groups.

The Mann-Whitney U and Wilcoxon signed-rank tests are powerful tools for analyzing data that does not meet the assumptions of parametric tests. You can make more informed statistical decisions by understanding their key differences and when to use them.

C) Correlational: Pearson’s r - Parametric:

Pearson's r is a statistical measure quantifying the linear correlation between two continuous variables. It is a parametric test (Salman & Aleem, 2024) that relies on specific assumptions about the population data, such as normality and homogeneity of variance. It assumes a linear relationship between the two variables. Both variables should be normally distributed. It requires data measured on an interval or ratio scale. The variance of one variable should be constant across all values of the other variable.

In interpreting Pearson's r, the Range Values range from -1 to +1. The closer the absolute value of r is to 1, the stronger the linear relationship. A positive linear relationship (as one variable increases, the other increases). A negative linear relationship (as one variable increases, the other decreases). r = 0: No linear relationship.

For Visual Representation: Pearson’s r is used to quantify the strength and direction of a linear relationship between two continuous variables and test the correlation's statistical significance. However, it is unsuitable for non-linear relationships, and outliers can significantly influence the correlation coefficient.

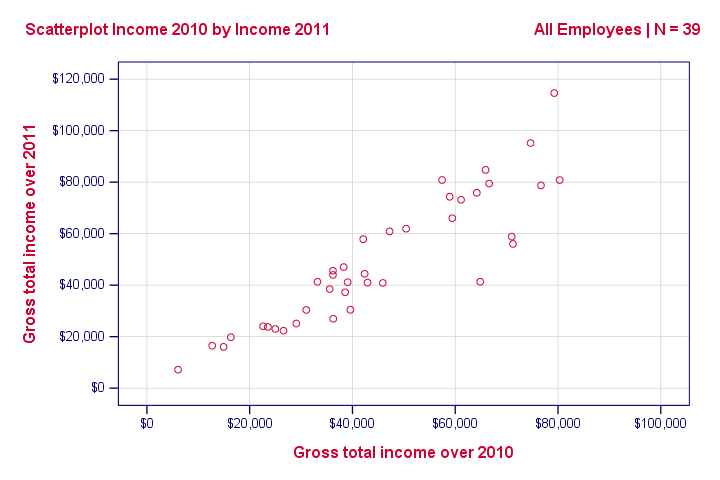
The alternatives to Pearson's r: If the data does not meet the assumptions of Pearson's r, it is advised to consider these alternatives: a) Spearman's Rank Correlation Coefficientis used for ordinal data or when the assumption of normality is violated. b) Kendall's Tau is used for ordinal data, mainly when many tied ranks exist. By understanding the characteristics and limitations of Pearson's r, you can effectively use it to analyze the relationship between variables in the research.

**The Pearson Correlation Visualization Fig 3.1**

A red and blue graph

Description automatically generated

**The Pearson Scatterplot Visualization Fig 3.2**



**Perfect, Strong, and Weak Correlation Figure 3.3**

A screenshot of a computer screen

Description automatically generated

**Key**

* **Figures 3.1 to 3.3** are culled from https://images.search.yahoo.com/yhs/search;\_ylt=AwrjecqyhTtn4ZUF4GV

wHmVH;\_ylu=Y29sbwNncTEEcG9zAzEEdnRpZAMEc2VjA3Nj?p=Example+of+Pearson+Correlation+visualization&type=Y243\_F881\_231651\_111324&hsimp=yhs-011&hspart=trp&grd=1&ei=UTF-8&fr=yhs-trp-011.

* **Figure 3.3** illustrates the Pearson Correlation Visualizations:
* Includes the perfect correlation, firm association, weak or no association.

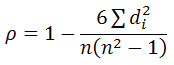
D) Correlational: Spearman’s Rank vs. Parametric Correlations:

A Quick Comparison: In Spearman's Rank Correlation, Non-parametric Do not rely on assumptions about population distribution (e.g., normality).   Ordinal Data is Suitable for ordinal data (ranked data).  The Monotonic Relationship measures the strength and direction of a monotonic relationship (either increasing or decreasing).   It is less sensitive to outliers compared to parametric tests.   It ranks the data for each variable and calculates the correlation between the ranks.

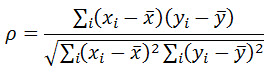
Parametric Correlations (e.g., Pearson's Correlation): There are parametric correlations when the data is usually distributed, and when the relationship between variables is linear, the data is interval or ratio. Parametric: Relies on specific assumptions about the population distribution (e.g., normality). It is suitable for interval or ratio data (continuous data). It measures the strength and direction of a linear relationship. It is more sensitive to outliers. Moreover, it directly calculates the correlation between the raw scores.

It is advised to use Spearman's Rank when data is not normally distributed. Moreover, when the relationship between variables is not linear but monotonic, and when the data is ordinal.

Spearman's Rank is a more flexible and robust correlation measure, especially when dealing with non-normal or ordinal data.  There are two methods to calculate Spearman's correlation depending on whether (1) your data. It does not have tied ranks; (2) your data has tied ranks. The formula for when there are no tied ranks is:



Where di = difference in paired ranks and *n* = number of cases. The formula to use when there are tied ranks is:



Where *i* = paired score. Extracts from (Laerd-Statistics, 2018, para. 11-12). Lund Research Ltd. https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical

-guide.php.

**4)** Navigate to OGS’s Practical Statistics for Social Research (PSSR) tool. Click on

“Example Datasets” and load the “Independent t-Test: Ethical Decision-Making.”

dataset. Scroll to “Step Three: Run Statistical Procedures” and click “t-Test”.

Copy and paste the output's contents into your assignment document. Read it carefully:

**Step Three: Run Statistical Procedures**  **Table 4.1**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| An Independent t-test (Student’s t-test or Welch’s Test) is a parametric procedure that compares the means of two independent groups. Sample sizes can be different, especially when using Welch’s test. It is used when data are typically distributed, and scales are interval or ratio. It focuses on whether the means of the two groups differ significantly.  **Results of t-Test Procedure**  **Table 1**  **T-Test Statistics**   |  |  | | --- | --- | | Measure | Value | | Group 1 (Religious Participants) Mean | 9.1625 | | Group 2 (Non-Religious Participants) Mean | 4.3630 | | Degrees of Freedom | 34.0000 | | t-Statistic | 3.8475 | | p-Value | 0.00050021 |   **Hypotheses**  **H0:** No statistically significant difference exists between Religious and Non-Religious Participants in the Ethical Decision-Making Scale (EDMS) scores.  **Ha**: A statistically significant difference exists in the Ethical Decision-Making Scale (EDMS) scores between Religious and Non-Religious Participants.  A two-tailed Student's t-test was applied to independent samples assuming equal variances to test the null hypothesis that the difference in means of the Ethical Decision-Making Scale (EDMS) between Religious and Non-Religious Participants was not equal to zero. The means for the groups Religious Participants and Non-Religious Participants were 9.1625 and 4.3630, respectively. With 34.0000 degrees of freedom, the t-statistic was 3.8475.  **The p-value** of 0.00050021 suggests a statistically significant difference between the groups' means at a 0.05 alpha level. The null hypothesis was rejected.  Post-Hoc Procedures  Table 2  T-Test Post Hoc Statistics   |  |  | | --- | --- | | Statistic | Value | | Bonferroni Correction Alpha | 0.050000 | | Cohen's d (effect size) | 1.290490 | | Power | 0.0000 | |

**Answer the following questions:**

**Qs. 4.1.1:** What might be good problem and purpose statements for this dataset? :

The problem is determining if there is a statistically significant difference in the mean scores of the Ethical Decision-Making Scale (EDMS) between Religious and Non-Religious Participants. This study aims to compare the mean scores of the EDMS between these two groups to identify any significant differences. There is a perceived difference in ethical decision-making between individuals who identify as religious and those who do not.

**Qs. 4.1.2**: What might be good research questions related to the hypotheses generated by the PSSR software?

The research question guiding this analysis is: Is there a statistically significant difference in the mean scores of the EDMS between Religious Participants and Non-Religious Participants?

**Qs 4.1.3:** What does the output tell you about comparing the two groups?

The mean score for Religious Participants is μ1=9.1625, while for Non-Religious Participants it is μ2=4.3630. The calculated t-statistic is t=3.8475 with df=34 degrees of freedom. The p-value associated with this t-statistic is p=0.00050021.

**Qs. 4.1.4:** How was the p-value used to test the hypotheses?

Since the p-value p=0.00050021 is less than the significance level α=0.05, we reject the null hypothesis H0. This indicates that there is a statistically significant difference between the means of the two groups.

**No. 4.5: Repeat steps 10-12 for** the “Correlational: Life Satisfaction Index” dataset. Note

that this is a correlational design with two continuous variables. What does the scatter plot tell you about the relationship between the two variables?

**“Perfect Correlational: Life Satisfaction Index” dataset. Table 4.5**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| An **Independent t-test** (Student’s t-test or Welch’s Test) is a parametric procedure that compares the means of two independent groups. Sample sizes can be different, especially when using Welch’s test. It is used when data are typically distributed, and scales are interval or ratio. It focuses on whether the means of the two groups differ significantly.  **Results of Linear Regression Procedure**  **Table 3**   |  |  | | --- | --- | | Measure | Value | | Slope | 1.239432 | | Intercept | 6.755420 | | R-squared | 0.133330 | | F-Stat. | 2.615309 | | p-Value | 0.12424038 |   **Linear Function**  **Table 4**   |  |  |  | | --- | --- | --- | | Variable | Symbol | Group | | Independent Var. | X | Hours of Weekly Religious Involvement | | Dependent Var. | f(X) | Life Satisfaction Index (LSI) |   f(x) = 1.24 \* X + 6.76  **Hypotheses**  H0: No statistically significant relationship exists between Hours of Weekly Religious Involvement and Life Satisfaction Index (LSI) as modeled by a regression function with a slope (coefficient) equal to zero (0).  Ha: A statistically significant relationship exists between Hours of Weekly Religious Involvement and Life Satisfaction Index (LSI) as modeled by a regression function with a slope (coefficient) approaching one (1).  **Findings**  A linear regression model was conducted to examine the extent to which the variability in the dependent variable f(x) or Y (Life et al. (LSI)) is predicted by the independent variable X (Hours of Weekly Religious Involvement).  The model explains a small portion of the response data's variability around its mean. The null hypothesis could not be rejected (p = 0.12424038).  **Scatterplot**   |  |  | | --- | --- | | 98.96  57.62 |  |   45.24 81.04 |

**4.5.1** What does the scatter plot tell you about the relationship between the two variables:

The scatter plot provides a visual representation of the relationship between Hours of Weekly Religious Involvement (X-axis) and Life Satisfaction Index (LSI) (Y-axis).

There appears to be a weak positive relationship between the two variables.

The scatter plot does not suggest a strong correlation between the two variables. The points are widely dispersed, indicating that the life satisfaction index for any given level of religious involvement is highly variable. The scatter plot supports the findings of the linear regression analysis.

**4.6 Repeat steps 10-12 for** both datasets but use the nonparametric equivalents of the statistical tests (Mann \* \* \* Whitney U and Spearman’s Rank, respectively).

**Mann-Whitney U and the Nonparametric procedure Table 4.6**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mann-Whitney U** is a nonparametric procedure that compares two independent groups. It does not require equal sample sizes. It is used when data is not normally distributed. The procedure assesses whether one group tends to have higher values than the other.  **Results of Mann-Whitney U Procedure**  **Table 5**  Mann Whitney U Statistics   |  |  | | --- | --- | | Measure | Value | | Group 1 (Hours of Weekly Religious Involvement) Mean | 1.7579 | | Group 2 (Life et al. (LSI)) Mean | 8.9342 | | Mann Whitney U | 12.0000 | | Z-Score | -4.9193 | | p-value | 0.00000087 |   **Hypotheses**  H0: No statistically significant difference exists in the distributions between Hours of Weekly Religious Involvement and Life Satisfaction Index (LSI).  Ha: A statistically significant difference exists in the distributions between Hours of Weekly Religious Involvement and Life Satisfaction Index (LSI).  A two-tailed Mann-Whitney U procedure was applied to the samples, assuming equal variances, to test the null hypothesis that the mean difference between the Hours of Weekly Religious Involvement and Life Satisfaction Index (LSI) was not equal to zero. The means for the group's Hours of Weekly Religious Involvement and Life Satisfaction Index (LSI) were 1.7579 and 8.9342, respectively.  The p-value of 0.00000087 suggests a statistically significant difference between the groups' distributions at a 0.05 alpha level. The null hypothesis was rejected.  **Post-Hoc Procedures**  **Table 6**  Mann-Whitney U Post Hoc Statistics   |  |  | | --- | --- | | Statistic | Value | | Bonferroni Correction Alpha | 0.050000 | | Cohen's d (effect size) | -2.378756 | | Power | 1.0000 | |

**Qs: 4.7**: For each dataset, click “Assumptions” under “Step Two: Run Descriptives

and Assumptions.” Based on the output of each dataset, should you use parametric or nonparametric procedures? Are the datasets normally distributed?

**Descriptives and Assumptions Based on Output Data Sets Table 4.7**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| In general, parametric tests are more powerful than nonparametric tests. However, they rely on assumptions about the distribution of the data. If these assumptions are not met, nonparametric tests may be more appropriate. If the results of the parametric and nonparametric tests are similar, it suggests that the assumptions of the parametric test were likely met. If the results differ, it may be necessary to consider the specific characteristics of the data and the assumptions of the tests to interpret the findings.    **Findings**  A linear regression model was conducted to examine the extent to which the variability in the dependent variable f(x) or Y (Spring) is predicted by the independent variable X (Fall). The model explains a small portion of the response data's variability around its mean. The null hypothesis could not be rejected (p = 0.37740680).  **Scatterplot**   |  |  | | --- | --- | | 98.96  57.62 |  |   45.24 81.04  **Point of Linear Function**  **Table 17**   |  |  |  | | --- | --- | --- | | Variable | Symbol | Group | | Independent Var. | X | Fall | | Dependent Var. | f(X) | Spring |   f(x) = -0.20 \* X + 89.39  **Hypotheses**  H0: No statistically significant relationship exists between Fall and Spring as modeled by a regression function with a slope (coefficient) equal to zero (0).  Ha: A statistically significant relationship exists between Fall and Spring as modeled by a regression function with a slope (coefficient) approaching one (1).  **Findings**  A linear regression model was conducted to examine the extent to which the variability in the dependent variable f(x) or Y (Spring) is predicted by the independent variable X (Fall).  The model explains a small portion of the response data's variability around its mean. The null hypothesis could not be rejected (p = 0.37740680). |

**4.7.1** Are the datasets usually distributed?

The data sets from facial visualization are scattered and not evenly distributed. However, to determine if the datasets are typically distributed, we can also use statistical tests like the Shapiro-Wilk test or visually inspect the data using histograms or Q-Q plots. Notably, normality is often assumed for parametric tests like linear regression, but it is not strictly required. The linear regression model can still be valid if the residuals (the differences between the observed and the predicted values) are typically distributed.

**5)** Finally, navigate OGS’s Practical Statistics for Social Research (PSSR) tool. Click

on “Example Datasets” and load the “Example: Perfect Correlation” dataset. Scroll to “Step Three: Run Statistical Procedures” and click “Linear Regression”. What does the scatterplot graph show you about the relationship between the X and Y variables?

**Table 5.1**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Results of Linear Regression Procedure**  Table 1   |  |  | | --- | --- | | Measure | Value | | Slope | 1.200000 | | Intercept | 0.250000 | | R-squared | 1.000000 | | F-Stat. | ∞ | | p-Value | 0.00000000 |   **Linear Function**  **Table 2**   |  |  |  | | --- | --- | --- | | Variable | Symbol | Group | | Independent Var. | X | X | | Dependent Var. | f(X) | Y |   f(x) = 1.20 \* X + 0.25  **Hypotheses**  H0: No statistically significant relationship exists between X and Y as modeled by a regression function with a slope (coefficient) equal to zero (0).  Ha: A statistically significant relationship exists between X and Y as modeled by a regression function with a slope (coefficient) approaching one (1).  **Findings**  A linear regression model was conducted to examine the extent to which the variability in the dependent variable f(x) or Y (Y) is predicted by the independent variable X (X).  The model explains all of the response data's variability around its mean. The null hypothesis was rejected (p = 0.00000000).  **Scatterplot**   |  |  | | --- | --- | | 30.25  1.45 |  |   1 25 |

**5.1** What does the scatterplot graph show you about the relationship between the X and Y variables?

The scatterplot ideally depicts a perfect linear relationship (Gallagher, 2024; Glaros, 2024) between the X and Y variables. This is indicated by the R-squared value of 1.000000, which means that the variation in X can explain 100% of the variation in Y. The data points would form a straight line in such a scenario, suggesting a strong positive correlation. Y also increases proportionally as X increases, following the linear function f(x) = 1.20 \* X + 0.25.

**6)** Summarize what you learned from conducting these statistical tests. Include a title page, well-developed introduction and conclusion paragraphs, a references page, and in-text APA-formatted support for your responses:

**Conclusion**

Getting perfect results in statistical surveys and experiments without errors is not admissible but can be minimized, and perhaps close to some data validity is crucial. In this study, one of the significant areas of learning is using the ubiquitous SPSS software to conduct several tests (Reichard, 2024) in linear regression and its procedures, as shown in Table 5.1, page 18. It includes the Perfect Correlation involving the Life Satisfaction Index dataset in Table 4.3, page 15. Others include the Mann-Whitney U in Table 4.6, page 16, and the Nonparametric procedure, outlining the “Descriptives and Assumptions Based on Output Data Sets in Table 4.7, page 17, in the SPSS software. Some of the scatterplots likely show a perfect linear relationship, with all data points falling directly on the regression line in Table 5.1, pg. 19. And some do not and are scattered in Table 4.7, pg. 17. This is consistent with the R-squared value of 1.00. In interpreting the Hypotheses, the Null Hypothesis (H0) (Reichard, 2024; Auger & Normand, 2024) study shows no significant linear relationship between X and Y shown in some of the SPSS model results. However, in the Alternative Hypothesis (Ha), There is a significant linear relationship between X and Y. It implies that the results have rejected the null hypothesis, supporting the alternative hypothesis. The linear regression model provides a highly accurate prediction of Y based on X. The relationship is strong, positive, and statistically significant. However, while the model may or may not be perfect in all cases, it is essential to remember that real-world data often has some degree of variability. It is always good practice to consider other factors and potential limitations when interpreting and applying the results. Linear regression remains a powerful statistical tool (Duan et al., 2024; Eka et al., 2024) used to model the relationship between a dependent variable and one or more independent variables. By fitting a linear equation to the data, one can make predictions, assess the strength of the relationship, and identify significant factors influencing the dependent variable. The results of the linear regression analysis provide valuable insights into the relationship between the variables. Linear regression is a versatile statistical technique (Adeleke et al., 2024; Grieve, 2024), and by understanding its assumptions and interpreting the results correctly (Waisapy, 2024; Tomaszewski et al., 2020), one can draw meaningful conclusions from the experimental data to ensure reliable inferences.

**Works Cited**

Adeleke, A. K., Montero, D. J. P., Olu-Lawal, K. A., & Olajiga, O. K. (2024). Statistical

Techniques in precision metrology, applications, and best practices. *Engineering Science & Technology Journal*, *5*(3), 888-900.

Ashton, J. C. (2013). Experimental power comes from influential theories, which is the real

problem in null hypothesis testing. *Nature Reviews Neuroscience*, *14*(8), 585–585.

Auger, V., & Normand, A. (2024). Data simulations for advancing psychological research:

Insights, preparations, and investigations. *International Journal of Psychology*.

Duan, R., Liang, C. J., Shaw, P. A., Tang, C. Y., & Chen, Y. (2024). Testing the missing at

Random assumption in generalized linear models in the presence of instrumental variables. *Scandinavian Journal of Statistics*, *51*(1), 334-354.

Eka, A. P. B., Bakri, A. A., & Yuliyani, L. (2024). Utilizing linear regression to forecast the

Stock price fluctuations of top-rated companies. *Jurnal Info Sains: Informatika dan Sains*, *14*(01), 551 559. https:ejournal.sea institute.or.id/index.php/Info Sains/article

/view/4041/3235.

EZ SPSS Tutorials (2024).- Simple Linear Regression in SPSS, Including Interpretation. Built

With Generated Press.

Gallagher, J. R. (2024). *Case Study Research in the Digital Age*. Taylor & Francis.

Glaros, A. G. (2024). Statistical significance, clinical importance, and effect sizes: Enhancing

understanding of a study's results. *Journal of Oral Rehabilitation*.

Grieve, A. P. (2024). Optimizing the Trade-Off Between Type I and Type II Errors: A Review

And Extensions. *arXiv preprint arXiv:2409.12081*.

Hadfield, C., Tyson, N., & Goodall, J. (2022). Master Class. Random Sampling Explained: What

Is Random Sampling?

Hatcher, L. (2013). *Advanced statistics in research: Reading, understanding, and writing up*

*data analysis results.*

Kelter, R. (2024). The Bayesian simulation study (BASIS) framework for simulation studies in

statistical and methodological research. *Biometrical Journal*, *66*(1), 2200095.

KIEU, Q. T., & MINH, H. (2024). Handbook for Professional Preparation of Research Papers &

Dissertations.

Kronthaler, F., & Zöllner, S. (2021). Data analysis with RStudio. *Data Analysis with RStudio*.

Kumar, A. (2024). Understanding Statistical Analysis: Techniques and Applications By

simplilearn.com/what-is-statistical-analysis-article.

Litt, S. (2024). A shortlist of the 10 Best Statistical Analysis Software Options. The CFO Club.

Financial Planning & Analysis.

MacFarland, T. W., Yates, J. M., MacFarland, T. W., & Yates, J. M. (2016). Mann–Whitney you

Test. *Introduction to nonparametric statistics for the biological sciences using R*, 103-

132.

Okoye, K. (2024). *R Programming: Statistical Data Analysis in Research*. Springer Nature.

Ose, S. O. (2016). Using Excel and Word to structure qualitative data. *Journal of Applied Social*

*Science*, *10*(2), 147-162.

Privitera, G. J. (2024). *Research methods for the behavioral sciences*. Sage Publications.

Pirani, S. (2024). Simplifying statistical Decision Making: A Research Scholar’s Guide to

parametric and Non-Parametric Methods. *International Journal of Multidisciplinary Research & Reviews*, *3*(03), 184-192.

Ravid, R. (2024). *Practical statistics for educators*. Rowman & Littlefield.

Reichard, J. (2024). Statistics as a Language: Overview of Statistical Test and Hypothesis

Testing. Video Tutorial.

Plodder, A., & Hamann, J. (2021). Practices of ethnographic research: Introduction to the special

issue. *Journal of Contemporary Ethnography*, *50*(1), 3-10.

Salman, H. M., & Aleem, D. N. (2024). An easy way to understand the Research Tools, Scales

And Parametric and non-parametric Tests. *Scales and Parametric & Non-Parametric Tests (October 14, 2024)*.

Shatz, I. (2024). Assumption-checking rather than (just) testing: The importance of visualization

and effect size in statistical diagnostics. *Behavior Research Methods*, *56*(2), 826–845.

Sileyew, K. J. (2019). *Research design and methodology* (Vol. 7). Cyberspace. https://www.

academia.edu/72232918/Research\_Design\_and\_Methodology.

Tomaszewski, L. E., Zarestky, J., & Gonzalez, E. (2020). Planning qualitative research: Design

And decision-making for new researchers. *International Journal of qualitative*

*methods*, *19*, 1609406920967174.

Waisapy, J. Y. (2024). Influence of Strategic Planning on Banking Performance via

Differentiation Strategy. *Mimico*, *1*(7), 65-77.

Waples, J. (2024). Simple Linear Regression: Everything You Need to Know. Data

Camp. https://www.datacamp.com/tutorial/simple-linear-regression.

Yang, H. (2013). The case for being automatic: Introducing the automatic linear modeling

(LINEAR) procedure in SPSS statistics. *Multiple Linear Regression Viewpoints*, *39*(2), 27-37.

Yashaswi, G. (2024). Discrete Vs. Continuous Data: Everything You Need To Know

Explore the differences between discrete and continuous data and their impact on

business decisions and customer insights.