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

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

**Assignment No. 1**

**Dr. Peter Abraham Airewele, DSL**

**Omega Graduate School**

**Professor**

**Dr. Sean Taladay, Ed.D**

 **November 4, 2024**

**Assignment #1: Discussion Questions**

Answer the following questions in short answer format and be prepared to discuss them with your classmates in the virtual residency or the discussion forum.

1. What is the purpose of assumptions tests, and what do they tell us about a given sample of data?
2. How are histograms and box-whisker plots helpful when evaluating a sample against a normal distribution?
3. How does normality help inform whether we can use parametric (generalizable to the target population) vs nonparametric procedures (applicable only to the sample)?
4. Determine when to use a t-test and Pearson’s r or their nonparametric equivalents to test hypotheses.
5. Navigate to OGS’s Practical Statistics for Social Research (PSSR) tool. Click on “Example Datasets” and load the “Dependent t-Test: Achievement Scores” dataset. Click on “Descriptives” and then on “Assumptions.” What do the histograms and box and whisker plots tell you about the normality of the samples? Click “Tools and Options” and “Generate Normal Distributions.” Re-run the “Descriptives” and “Assumptions.” How are the normal distributions different from the original samples?

**Introduction**

This study covers some statistical principles and practices of assumption testing to determine the efficacy in evaluating and validating a given data set, emphasizing the trend in the Groups’ dependent and independent variability. The similarity between the histograms and box-whisker plots is explored to establish why they are helpful when evaluating a sample against a normal distribution. Moreover, the t-tests used to compare the means of two groups (Fiandini et al., 2024; Li et al., 2024; Parekh, 2024) usually require distributed data. The other areas of study are the parametric and non-parametric procedures, which are used when there is a large or specific sample size or when the sample meets certain assumptions about normality. The nonparametric test is used in the specific samples analyzed, not the wider population. Other areas of study are skewness, which measures data asymmetry around the mean, and Excess kurtosis, which assesses tail heaviness relative to a normal distribution, indicating more or fewer outliers. The Kolmogorov-Smirnov statistic tests are designed to match a specified distribution, with larger values suggesting more significant divergence. The Shapiro-Wilk statistic tests normality are used where significant results imply a non-normal distribution (A’aqoulah et al., 2024; Reichard, 2024) and are crucial. The Levene’s Test and the F-Test for Equal Variances check for homogeneity across groups, essential for analyses assuming equal variances. Together, these tests inform the suitability of data for further parametric statistical procedures (Reichard, 2024). These statistical tests are crucial (de Souza et al., 2024; Yagin et al., 2024) in navigating multiple statistical inductive or deductive, descriptive or quantitative output findings.

1. What is the purpose of assumptions tests, and what do they tell us about a given sample of data?

 Assumptions tests are used to validate whether a given data sample meets the necessary conditions and the assumptions required to accurately apply a specific statistical test (Shatz, I. (2024; Komatsu et al., 2024), essentially telling us if the conclusions drawn from the analysis are reliable and valid; if the assumptions are not met, the results of the statistical test may be inaccurate or misleading.  The tabular tables and charts clearly show the data dynamics and how they are distributed in linear or non-linear progressions.

The reminder is that assumptions are checked to ensure that the chosen statistical test is appropriate for the data and that the results can be confidently interpreted. Moreover, If assumptions are not met, it may indicate problems with data collection, data distribution, or the chosen statistical method, prompting the need to address these issues or select a different analysis technique.

**2)** How are histograms and box-whisker plots helpful when evaluating a sample against a normal distribution?

Histograms and box-whisker plots are helpful when assessing a sample alongside a normal distribution (Komatsu et al., 2024; Sandfeld, 2023) because a histogram visually shows the overall shape of the data distribution (Magrab, 2022; Geng et al., 2024; Shatz, 2024) allowing you to see if it is similar to a bell curve (symmetrical and unimodal) characteristic of a normal distribution. By looking at the height of the bars, you observe the density and can assess the concentration of data points in different ranges mentioned earlier.

The box-whisker plot highlights prospective skewness by showing the position of the median relative to the quartiles. It can also recognize outliers, which could indicate a deviation from normality. You quickly detect the skewness if the median line within the box is not centered; it suggests a skewed distribution. Points outside the whiskers are considered potential outliers, indicating non-normality.  The relative lengths of the "box" (interquartile range) on either side of the median can also provide insight into potential skewness. Some argue that while visual assessments are helpful, it is important to use statistical tests like the Shapiro-Wilk test to formally evaluate normality, especially when making critical decisions based on the data.

**The box and whisker plot examples:**

**The Assumption** is: that seven students get the following scores in 2 subjects:

 Figure 2.1

| **Student** | **Ann** | **Jacob** | **Susan** | **Rose** | **Tom** | **Peter** | **Kate** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Scores in Maths | 35 | 45 | 56 | 66 | 71 | 78 | 91 |
| Scores in Physics | 30 | 45 | 68 | 70 | 78 | 83 | 85 |

A quick summary of the values for Box and Whisker chart are: Figure 2.2

| **Maths** | **Physics** |
| --- | --- |
| Median= 66 | Median= 70 |
|  |  |
| Lower quartile= 42 | Lower quartile= 45 |
|  |  |
| Upper quartile= 78 | Upper quartile= 83 |
|  |  |
| Minimum value= 35 | Minimum value= 30 |
|  |  |
| Maximum value= 91 | Maximum value= 85 |

|  |
| --- |
| **Test Scores of 7 Students in 2 Subjects** |
| 100 |  |  91 |  85 | Third Quartile (Physics) |
| 80 |  |  |  | Third Quartile (Maths) |
| 60 |  |  |  |  |
| 40 |  |  |  | First Quartile Physics |
| 20 |  |  35 |  30 | First Quartile (Maths) |
|  |  | Maths  | Physics |  |

 Figure 2.3

 Figure 2.4



Figure 2.5



**Key**

* **Figures 2.1 to 2.3 are Box & Whisker Charts.** They are culled from: Source: Box and whisker plots / https://www.fusioncharts.com/resources/chart-primers/box-and-whisker-chart.
* The Box and Whisker charts are simple and can be analyzed and illustrated visually or digitally.
* **Figures 2.4 and 2.5** are culled from Quartile: Definition, Finding, and Using- Statistics By Jim Frost.

The simple method for finding quartiles is to list the values in the dataset in numeric order. Then, find the three values that split your data into quarters, as shown above and below. Quartile, Q1 is 20, Q2 is 29, and Q3 is 39, as shown below.

Quartiles are three values that split your dataset into quarters. These values are the following:

* Q1 First quartile: 25% of the data are below this value.
* Q2: Second quartile / Median: This value splits the data in half.
* Q3 Third quartile: 25% of the data exceeds this value (Frost, 2024).

3) How does normality help inform whether we can use parametric (generalizable to the target population) vs nonparametric procedures (applicable only to the sample)?

Parametric procedures are used when there is a large sample size or when the sample meets certain assumptions about normality. Normality helps determine whether to use parametric or nonparametric statistical procedures by indicating whether the data is likely drawn from a population with a normal distribution, a key assumption for parametric tests. However, suppose the data is not normally distributed. In that case, nonparametric tests, which make no such assumption, should be used instead, meaning the results will only be generalizable to the specific sample analyzed, not the wider population.

Parametric tests assume the data is normally distributed (or follows a known distribution) and can be used to make inferences about population parameters like

Mean and standard deviation. They allow for generalization to the broader population if the normality assumption holds.

Nonparametric tests do not require the assumption of normality, making them suitable for data with unknown distributions or skewed data (Rani, 2024; Okoye & Hosseini, 2024). However, the results are typically only applicable to the sample studied. To assess these tests, you examine the histograms, or Q-Q plots, to see if the data appears roughly symmetrical and bell-shaped. Performing normality tests, such as the Shapiro-Wilk test, will require assessing the likelihood of the data being normally distributed statistically.

 **4)** Determine when to use a t-test and Pearson’s r or their nonparametric equivalents to test hypotheses.

A t-test is used to compare the means of two groups, and the data is usually distributed while using Pearson's- r when you want to assess the linear relationship between two continuous variables that are also normally distributed. For example, suppose the data does not meet these normality assumptions. In that case, nonparametric equivalents like the Mann-Whitney U test (for comparing groups) or Spearman's rank correlation (for assessing relationships) should be used instead.  The two seem to show some similarities when utilized.

The T-test compares the means of two groups and requires typically distributed data. The Nonparametric alternative is the Mann-Whitney U test.

Pearson's r measures the strength of a linear relationship between two continuous variables. Both variables must be normally distributed.

Moreover, the nonparametric tests are used when the data is not normally distributed (Staňková & Střelec, 2024), when the data is ordinal (ranked), or a small sample size (mentioned earlier). Statistical procedures complemented these areas of study, the parametric (Pearson’s correlation) and the non-parametric (Spearman’s Rank and Kendall Tau-b procedures).

**5) Navigate** to OGS’s Practical Statistics for Social Research (PSSR) tool. Click on “Example Datasets” and load the “Dependent t-Test: Achievement Scores” dataset. Click on “Descriptives” and then on “Assumptions.” What do the histograms and box and whisker plots tell you about the normality of the samples? Click “Tools and Options” and “Generate Normal Distributions.” Re-run the “Descriptives” and “Assumptions.” How are the normal distributions different from the original samples?

What the histograms and box and whisker plots tell about the normality of the samples:

* They show two Groups, 1 and 2, with continuous variables, with the mean values of zero distributions.
* The Pearson’s r-coefficients are NAN, not of values. They show the direction of a linear relationship between two continuous variables.
* The p-value is NAN, not of values. It is an alternative to rejection points to provide the minor significance level at which the null hypothesis may be rejected. A smaller p-value means more substantial evidence in favor of the alternative hypothesis.
* There is a mix of Ho, nonstatistical significance, and Ha statistically significant hypotheses.
* **In Step 3, The Run Statistical Procedures** **show:**

In the Ho hypothesis, No statistically significant relationship exists between Group 1 and Group 2.

In the Ha hypothesis, A statistically significant relationship exists between Group 1 and Group 2.

1. Quasi-experimental / Causal-Comparative Designs, namely:
2. the Parametric (Pearson’s correlation).
3. And the Non-Parametric (Spearman’s Rank and Kendall Tau-b).
4. Summary of Descriptive Assumption Tests: See details below:

Assumptions Tests in statistics evaluate whether data characteristics meet the conditions required for specific statistical tests.

Skewness measures data asymmetry around the mean; values far from zero suggest significant skew. Excess kurtosis assesses tail heaviness relative to a normal distribution, indicating more or fewer outliers.

The Kolmogorov-Smirnov statistic tests if a sample matches a specified distribution, with larger values suggesting more significant divergence.

The Shapiro-Wilk statistic tests normality, where significant results imply a non-normal distribution.

Levene’s Test and the F-Test for Equal Variances check for homogeneity across groups, essential for analyses assuming equal variances. Together, these tests inform the suitability of data for further parametric statistical procedures:

1. **In Step Four, the Review Results and output show** two Groups 1 and 2 of variables, the Pearson’s Correlation Coefficient of (r) as NAN, and the P-value as a NAN (Not a number) shows the Ho: Null values, and Ha: Alternative value relationships in Table # 8:

**Step Four: Review Results:**

**Output**

**Pearson’s Correlation Coefficient** is a parametric procedure to measure the strength and direction of a linear relationship between two continuous variables. It assumes that both variables are normally distributed. Interpretation is based on the correlation coefficient, ranging from -1 to +1, indicating negative or positive linear relationships.

**Results of Statistical Procedure**

**Table 8**

|  |  |
| --- | --- |
| **Measure** | **Value** |
| Group 1 (Group 1) Mean | 0.0000 |
| Group 2 (Group 2) Mean | 0.0000 |
| Pearson's Correlation Coefficient (r) | NaN |
| p-value | NaN |

**Hypotheses**

Ho: No statistically significant relationship exists between Group 1 and Group 2.

Ha: A statistically significant relationship exists between Group 1 and Group 2.

**Findings**

The r-value of NaN indicates a sizeable positive effect between Group 1 and Group 2.

The probability that the relationship between Group 1 and Group 2 was statistically significant at

a 95% confidence level (p = NaN). The null hypothesis could not be rejected.

**Scatterplot**

|  |  |
| --- | --- |
| Y000 |   |
| X |  0 |

**The Scatter plot** shows “0” values on the X and Y axis.

**Attribution**

Statistical procedures were conducted using [PSSR (Practical Statistics for Social Research)](https://stats.ogs.edu/), statistical analysis software developed by Joshua D. Reichard for [Omega Graduate School](https://ogs.edu/) based on the [jStat](https://jstat.github.io/) library.

**Reference**

Reichard, J. (2024). *Practical Statistics for Social Research (PSSR)*. Omega Graduate School. <https://stats.ogs.edu/>.

**Attribution**

Statistical procedures were conducted using [PSSR (Practical Statistics for Social Research)](https://stats.ogs.edu/), statistical analysis software developed by Joshua D. Reichard for [Omega Graduate School](https://ogs.edu/) based on the [jStat](https://jstat.github.io/) library.

 **Conclusion**

 Given the crucial relevance of groups’ dependent and independent variable data in descriptive, inductive, and quantitative deductive experiments (Komatsu et al., 2024), the assumption and the t-test, the parametric and non-parametric statistics have become indispensable in relative statistical research studies. This study shows that histograms and box-whisker plots are mostly embraced in comparative statistical experiments in academia, marketing, income distribution, and employment demographics because they are more straightforward to comprehend, assess, and evaluate by organizations, customers, and corporate clientele. Moreover, they are helpful when assessing a sample alongside a normal distribution because a histogram visually shows the overall shape of the data distribution, allowing you to see if it is similar to a bell curve, symmetrical and unimodal characteristic of a normal distribution.

**Works Cited**

A’aqoulah, A., Albalas, S., Albalas, M., Alherbish, R. A., & Innab, N. (2024). Comparative

Analysis and Spatial Distribution of the Primary Health Care Centers and Health Manpower Across Saudi Arabia Using Shapiro–Wilk Test. *Journal of Multidisciplinary Healthcare*, 4851-4861.

de Souza Junior, R. S., & Borges, E. M. (2024). Teaching statistics and chemometrics using an

Open source, free, and graphical user interface software.

Fiandini, M., Nandiyanto, A. B. D., Al Husaeni, D. F., Al Husaeni, D. N., & Mushiban, M.

(2024). How to calculate statistics for significant difference tests using SPSS: Understanding students' comprehension of steam engines as power plants. *Indonesian Journal of Science and Technology*, *9*(1), 45-108.

Frost, J. (2024). Quartile: Definition, Finding, and Using - Statistics by Jim.

Geng, Y., Li, Q., Yang, G., & Qiu, W. (2024). Basics of Machine Learning. In *Practical*

*Machine Learning Illustrated with KNIME* (pp. 15-43). Singapore: Springer Nature Singapore.

Jideani, V. A., Jideani, A. I., Zheleva, I., Todorov, T., Bozkurt, H., Göğüş, F., ... & Muntean, E.

(2024). Basic Statistics for Data Collection and Treatment. In *Statistics in Food and Biotechnology: From Theory to Practical Applications* (pp. 1-154). Cham: Springer International Publishing.

Komatsu, K., Murata, S., Stylianides, A. J., & Stylianides, G. J. (2024). Introducing Students to

The Role of Assumptions in Mathematical Activity. *Cognition and Instruction*, *42*(2), 327-357.

Li, Z., Feng, R., Wang, Y., & Wang, L. (2017). Experimental study on the effect of dents

Induced by the impact on the fatigue life of the 2024-T3 aluminum alloy plate. *Engineering Structures*, *137*, 236-244.

Magrab, E. B. (2022). Descriptive Statistics and Discrete Probability Distributions. In

*Engineering Statistics: An Introduction* (pp. 1-27). Cham: Springer International Publishing.

Parekh, R. (2024). Comparison Analysis of Construction Costs according to LEED and non-LEED Certified Educational Buildings. *Available at SSRN 4924703*.

Okoye, K., & Hosseini, S. (2024). Choosing Between Parametric and Non-parametric Tests in

Statistical Data Analysis. In *R Programming: Statistical Data Analysis in Research* (pp. 87–98). Singapore: Springer Nature Singapore.

Rani, S. U., Kishore, M. R., Rani, N., & Sunanda, N. (2024). Quantifying the Dispersion of

Rainfall through the Box Whisker Technique.

Reichard, J. (2024). *Practical Statistics for Social Research (PSSR)*. Omega Graduate School.

https://stats.ogs.edu/.

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.

Sandfeld, S. (2023). Exploratory Data Analysis. In *Materials Data Science: Introduction to Data*

*Mining, Machine Learning, and Data-Driven Predictions for Materials Science and Engineering* (pp. 179-206). Cham: Springer International Publishing.

Staňková, M., & Střelec, L. (2024). Efficiency Evaluation in Practice: a Comparison of

Parametric and Non-parametric Approaches: Accepted-May 2024. *REVSTAT-Statistical Journal*.

Yagin, F. H., Yagin, B., & Pinar, A. (2024). Normality Distributions Commonly Used in Sport

And Health Sciences. *Journal of Exercise Science & Physical Activity Reviews*, *2*(1), 124-131.