**LAB 2: INFERENTIAL STATISTICS FOR A DATASET**

**Prelab Questions/Exercises**

**1. Define inferential statistics and provide an example.**  
Inferential statistics involve using data from a sample to make generalizations or predictions about a larger population. It allows researchers to test hypotheses and estimate population parameters.  
*Example*: Suppose you survey 500 people to determine the average monthly expenditure on groceries in a city. Using inferential statistics, you can estimate the average monthly expenditure for the entire city's population.

**2. What is the difference between a t-test and a z-test?**

* **T-test**: Used when the sample size is small (usually n<30n < 30n<30) or when the population standard deviation is unknown. It compares the means of two groups or a sample mean to a population mean.
* **Z-test**: Used when the sample size is large (n≥30n \geq 30n≥30) and the population standard deviation is known. It is also used for comparing sample and population means when the data is approximately normally distributed.

**3. Why are p-values important in inferential statistics?**  
P-values help determine the significance of results in hypothesis testing. They measure the probability of observing a test statistic as extreme as the one calculated, assuming the null hypothesis is true.

* A small p-value (e.g., p<0.05p < 0.05p<0.05) indicates strong evidence against the null hypothesis, leading to its rejection.
* A large p-value (p>0.05p > 0.05p>0.05) suggests insufficient evidence to reject the null hypothesis.

**4. Explain the concept of confidence intervals.**  
A confidence interval is a range of values within which the true population parameter is likely to fall, with a certain level of confidence (e.g., 95%). It provides an estimate of the parameter along with an associated level of uncertainty.  
*Example*: If the mean height of a sample is 170 cm with a 95% confidence interval of 165–175 cm, we are 95% confident that the true mean height of the population lies between 165 and 175 cm.

**5. How can outliers impact hypothesis testing?**  
Outliers can significantly affect hypothesis testing because they can distort measures like the mean and standard deviation, leading to incorrect conclusions. For instance:

* They may increase the variability, reducing the statistical power of a test.
* They can lead to Type I (false positive) or Type II (false negative) errors by exaggerating or masking true effects.  
  Proper outlier detection and treatment are essential to maintain the accuracy of hypothesis tests.

**In-Lab**

* **Objective**: Apply inferential statistics to analyze and interpret data using hypothesis testing and confidence intervals.
* **Resources**: Python (Jupyter Notebook), SciPy library, and dataset (feedback\_data.csv) with customer ratings for two products.

<https://colab.research.google.com/drive/176Lrx_WAgtk-fAmd7OOwP3Q9T19sCVSk?usp=sharing>

**Postlab Questions**

**1. What does the p-value indicate in hypothesis testing?**  
The p-value measures the probability of obtaining test results as extreme as the observed results, assuming the null hypothesis is true. It indicates the strength of evidence against the null hypothesis:

* A small p-value (e.g., p<0.05p < 0.05p<0.05) suggests strong evidence to reject the null hypothesis.
* A large p-value suggests insufficient evidence to reject the null hypothesis.

**2. How can the t-statistic help in comparing two groups?**  
The t-statistic measures the difference between the sample means of two groups relative to the variability of the data. It helps determine whether the observed differences are statistically significant. By comparing the calculated t-statistic to a critical value (or using the p-value), we can decide whether to reject the null hypothesis.

**3. Explain why confidence intervals are preferred over point estimates.**  
Confidence intervals provide a range of values within which the true population parameter is likely to fall, along with a specified level of confidence (e.g., 95%). They are preferred over point estimates because:

* They quantify the uncertainty in the estimate.
* They offer more information than a single point estimate by showing the range of plausible values.

**4. If the p-value is 0.03, what does it mean in the context of hypothesis testing?**  
A p-value of 0.03 means there is a 3% probability of obtaining results as extreme as the observed ones if the null hypothesis is true. In many contexts, if the significance level (α\alphaα) is set to 0.05, the null hypothesis would be rejected, indicating that the results are statistically significant.

**5. Suggest improvements to handle unequal variances between the two groups.**  
To handle unequal variances, you can:

* Use **Welch’s t-test**, which does not assume equal variances.
* Transform the data (e.g., logarithmic transformation) to stabilize variances.
* Apply non-parametric tests such as the Mann-Whitney U test, which are less sensitive to unequal variances.
* Increase sample size to reduce variability in estimates.