**Quiz/Test #3 (Applied Stats) – Due Nov 26, 2024.**

**(Each question worth 10 points)**

**Show all work! State any resources or APPS you use in your solutions.**

1. Medical Malpractice. In a study of 1228 randomly selected medical malpractice lawsuits, it was found that 856 of them were dropped or dismissed. Construct a 95% confidence interval for the proportion of medical malpractice lawsuits that are dropped or dismissed.

**Solution:** Given:

* Sample size (n) = 1228
* Number of cases dropped/dismissed (x) = 856

Step 1: Calculate the sample proportion (p):

p=(x/n)

p= (856/1228) =0.696

So, the sample proportion (p) is approximately **0.696** or **69.6%**. This means that about 69.6% of the medical malpractice lawsuits in the sample were dropped or dismissed.

Step 2: Calculate the standard error (SE) for the proportion:

A black background with white text

Description automatically generated

Step 3: Find the z-score for a 95% confidence level (z = 1.96).

Step 4: Construct the confidence interval:

A black screen with white text

Description automatically generated

**95% Confidence Interval**: (0.6706, 0.7228)

Interpretation: We are 95% confident that the true proportion of medical malpractice lawsuits that are dropped or dismissed lies between 67.06% and 72.28%.

A screenshot of a computer

Description automatically generated

1. A Quest Dynamics analysis of 10 million drug tests revealed that 4.2% of them tested positive for illegal drugs. Construct a 95% confidence interval estimate AND a 99% confidence interval estimate of the positive test rate. Compare the results.

**Solution:** Given:

* Sample size (n) = 10,000,000
* Proportion of positive tests (p) = 0.042

**95% Confidence Interval**

Step 1: Calculate the standard error (SE):

Step 2: Find the z-score for a 95% confidence level (z = 1.96).

Step 3: Construct the confidence interval:

**99% Confidence Interval**

Step 1: Find the z-score for a 99% confidence level (z = 2.576).

Step 2: Construct the confidence interval:

**Comparison**:

* **95% Confidence Interval**: (0.0408, 0.0432)
* **99% Confidence Interval**: (0.0404, 0.0436)

Interpretation: The 99% confidence interval is wider than the 95% interval, reflecting greater confidence in capturing the true proportion at the cost of precision. The 95% interval provides a narrower range with slightly less certainty

1. In a test of weight loss programs, 40 adults used the Atkins weight loss program. After 12 months, their mean weight loss was found to be 2.1 lbs, with a standard deviation of 4.8 lbs. Construct a 90% confidence interval estimate of the mean weight loss for all such subjects. Does the Atkins program appear to be effective? Does it appear to be practical?

Solution: Given:

* Sample size (n) = 40
* Sample mean (\bar{x}) = 2.1 lbs.
* Sample standard deviation (s) = 4.8 lbs.

Step 1: Find the t-score for a 90% confidence level with . (Approximate value:)

Step 2: Calculate the standard error (SE):

Step 3: Construct the confidence interval:

*Interpretation*: We are 90% confident that the true mean weight loss for all such subjects lies between 0.821 lbs. and 3.379 lbs.

*Effectiveness*: The weight loss is modest and may not be substantial enough to be considered highly effective.

*Practicality*: With a mean loss of only 2.1 lbs., the program might not be practical for significant weight loss goals.

1. Find the sample size needed to estimate the mean age of movie patrons, given that we want 98% confidence that the sample mean is within 1.5 years of the population mean. Assume that the standard deviation of the population is 19.6 years. Do you think the sample could be obtained from one movie at one theater?

**Solution:** Given:

* Desired margin of error (E) = 1.5 years
* Standard deviation (σ) = 19.6 years
* Confidence level = 98% (z = 2.33)

Step 1: Use the formula for sample size (n):

**Answer**: A sample size of approximately 930 is needed.

Feasibility: Obtaining such a large sample size from one movie at one theater may not be practical. A broader sampling strategy involving multiple theaters and showtimes would likely be necessary.

Resources Used: Statistical sample size formulas and z-table.

1. Claim: Most adults know that a light year is a measure of distance.

Sample data: A Pew Research Center survey of 3278 adults showed that 72% knew that a light year is a measure of distance. Write the NULL and ALTERNATIVE hypothesis in symbolic form.

**Null Hypothesis (H₀)**: (50% or fewer adults know that a light year is a measure of distance)

**Alternative Hypothesis (H₁)**: (More than 50% of adults know that a light year is a measure of distance)

1. Original Claim: More than 58% of adults would erase all of their personal information online if they could. The hypothesis test results in a p-value of 0.3257. Assume a significance level of alpha = 0.05.

**p-value = 0.3257**

**Significance level (α\alphaα) = 0.05**

1. State a conclusion about the NULL hypothesis (REJECT or FAIL to REJECT)

**Hypothesis**: Since the p-value (0.3257) is greater than the significance level (α=0.05\alpha = 0.05α=0.05), we **fail to reject the null hypothesis**.

1. State the final conclusion about the NULL hypothesis in non-technical terms. (Use the table to below to construct your statement)

There is not enough evidence to support the claim that more than 58% of adults would erase all their personal information online if they could.

A diagram of a flowchart

Description automatically generated

1. Assume a linear regression model has been constructed from some dataset. The graph below shows a plot of the residuals. Comment on if any of the SLR assumptions (as documented in section 8.2 of your zyBook textbook) have been violated or otherwise look to be suspect. Address only the following assumptions:

A graph with dots and numbers

Description automatically generatedA graph of a function

Description automatically generated

The residual plot suggests violations of the assumptions of linearity, homoscedasticity, and independence.

The Q-Q plot indicates that normality is mostly satisfied but could be improved at the extreme values.

1. Might it be possible to use the method Simple Linear Regression on the following data? Assume we have one predictor and one response variable. If your answer is ‘yes’, please provide a sketch of possible data preprocessing step(s).

A graph with numbers and points

Description automatically generated

The scatter plot shows a non-linear pattern, suggesting a potential non-linear relationship between the variables. The variance appears to increase as the value of x increases, indicating heteroscedasticity, which violates one of the assumptions of Simple Linear Regression (SLR).

Recommendations for Data Preprocessing:

1. Transforming the Data: Consider using a logarithmic, square root, or power transformation on the response variable y to stabilize variance and make the relationship more linear.
2. Polynomial Regression: If transformations do not yield a linear relationship, using polynomial regression (e.g., including x2x^2x2 as an additional term) might be more suitable.
3. Splitting into Segments: If there are clear breakpoints, consider segmenting the data and fitting separate models for each segment.
4. Weighted Regression: If heteroscedasticity persists, apply a weighted regression approach where weights are inversely proportional to the variance of y for better handling varying spread.

Would you like additional information on any of these methods or guidance on implementing them?

1. Using the **mtcars** dataset in your textbook (zyBook). Explain the process you would follow in order to determine which SINGLE predictor variable explains MPG the best? Which predictor variable is best? Justify your answer using the diagnostic results that come from building linear regression models.
2. Using the results from 9), what is the second best variable to add to the model? Justify your results by using the diagnostic information produced during the modeling effort. Include any other metrics you would like to use to justify your answer. Pay attention to the

adjusted -R^2 value, and watch for any multi-collinearity between the predictor value you selected in 9) and whatever you plan to add as the second predictor value.

**Worth 10 extra points:**

**EXTRA CREDIT: College classes do not teach you everything. You are required to learn things on your own. Sometimes problems cannot be handled in textbook fashion or with well-defined distributions.**

Assume we have a ***very*** small sample of objects from the population. The weights for the 5 objects in our small sample are the following (in Kg):

120, 140, 80, 100, 240

Construct a 90% confidence interval for the mean weight of the population using the method of bootstrapping.

***Hint: ChatGPT … “explain the use of bootstrapping to estimate the mean of a population from a sample size with a very small number of objects”***

Final question, multiple choice, worth 0 points and entirely optional (even ignorable)!

Who is the “mostest bestest” stats teacher in the universe?

1. Greg MacDonald

😊 😊 😊

P.S. We should have some fun with serious things every now and then!