heaven sta130 proposal

November 4, 2024

research question1: if there exists a relationship between lonelyness index and social meida time per day
ho:there is no relationship between lonelyness index -and social media time per day
h1:there exists a relationship between lonelyness index and social meida time per day
independent variable:social media time per day(quantitive) measures how long a preson use social media per day
dependent variable: lonelyness ——index(quantitive) a index measures to what extent a preson is lonely

method chi-square test reason The chi-square test for independence is designed to assess whether there is an association between two categorical variables. It evaluates if the observed frequencies in a contingency table differ significantly from what would be expected if the variables were independent. This is particularly valuable for examining relationships in fields like social sciences, health studies, and marketing.

visualization: bar plot:By displaying bars of expected versus observed counts, bar plots allow users to visualize discrepancies for each category. heat map:Each cell color intensity represents the frequency count or proportion, making it easy to observe where high or low values concentrate within the dataset.

analysis Independence of Observations: Each data point is collected independently

Expected Frequency Requirements: each cell in the contingency table (representing categories of loneliness index against categories of social media usage time) should have an expected frequency of at least 5.

Categorical Variables: Both variables in this test—loneliness index and daily social media usage timeare treated as categorical variables, Random Sampling: The data was randomlly sampled The chi-square test results indicate a strong association between the two variables—loneliness index and daily social media usage timebased on the following statistical findings:

check the results:

P-value: The p-value of 4.97×10 ……-77 is far below the typical significance level (e.g., 0.05). Such an extremely low p-value indicates that the likelihood of observing this data if the null hypothesis were true

Degrees of Freedom: With 20 degrees of freedom

Expected Frequency Table: The expected frequency table shows that each cell has a frequency above 5, which satisfies the chi-square test assumption for expected frequencies. This validation of assumptions ensures that the test results are reliable and accurate.

research question2 if the sample mean is equal to a hypothesized population mean (H) for the sampling distribution of life satisfaction the sample mean is equal to a hypothesized population mean (H1) for the sampling distribution of life satisfaction the sample mean is not equal to a hypothesized population mean variable life satisfaction, reflects people's degree of satisfaction of their lifes after the pandemic

method Confidence Interval: Sampling distributions are essential for constructing confidence intervals. Knowing the variability in sample statistics (like the standard error) allows us to define intervals around a sample estimate within which the population parameter is likely to fall, adding reliability to our estimates.

sampling distribution Sampling distributions allow us to estimate population parameters, such as the mean, with greater accuracy. By analyzing the behavior of sample statistics (like sample means) over repeated samples, we gain a clearer picture of where the true population parameter likely lies.

visualization bar plot: straightforward and widely understood, making them an accessible choice for visualizing data. The clear visual differences between bar heights make it easy to interpret key patterns or deviations from expected values

hypothesis Independence of Observations: Each observation in the sample is collected independently, meaning that the value of one observation does not affect another. This assumption holds, as the data was collected randomly without interference among participants.

Normality of the Sampling Distribution: For the test statistic, the sampling distribution is approximately normal. If the sample size is sufficiently large (e.g., 30), the Central Limit Theorem guarantees that the sampling distribution of the mean (or other statistic) will approximate normality, regardless of the population's distribution. Given our sample size, this assumption is satisfied.

Sufficient Sample Size: The sample size is large enough to provide reliable results, reducing the impact of sampling variability. This large sample size alsocontributes to the normality of the sampling distribution, supporting the robustness of the test.

Random Sampling: The data was collected through a random sampling process, ensuring that the sample is representative of the population and that the results can be generalized to a broader context.

analysis The rejection of the null hypothesis because 5 is not within the confidence interval has significant implications in the context of our analysis.

research question3 use complex linear regression to fit and forecast life satisfaction independent variable COVID_vaccinated if you have received a COVID-19 Vaccine LONELY_direct During the PAST WEEK, if you have felt lonely LONELY_others_aware:Generally speaking, if you think others are aware of the extent to which you feel lonely or connected CONNECTION_social_num_close_friends:how many close friends does each of the investors has

method 1. Correlation Analysis Correlation analysis helps me understand the relationships between independent variables (predictors) and the dependent variable (outcome).

2. Visualization Using Bar Charts Using bar charts allows me to visually assess the relationships and distributions within the data, especially for categorical or ordinal predictors.

3. Data Splitting Splitting the dataset into training and testing subsets is essential to evaluate how well the model generalizes to new, unseen data.

4. Linear Regression Multiple linear regression is the core method used to quantify the relationship between the dependent variable and multiple predictors.
5. Multicollinearity Detection Multicollinearity detection, often conducted using the Variance Inflation Factor (VIF), helps me identify instances where predictors are highly correlated with each other

hypothesis Linearity: The relationship between each independent variable and the dependent variable should be linear.

Independence of Errors: The residuals (errors) should be independent of each other.

Homoscedasticity: This assumption states that the residuals should have constant variance across all levels of the independent variables.

Normality of Errors: The residuals should be approximately normally distributed, especially when inference (such as hypothesis testing and confidence intervals) is required.

No Multicollinearity: Independent variables should not be highly correlated with each other, as high multicollinearity can make it challenging to assess the individual effect of each predictor.

Mean of Residuals is Zero: The average of the residuals should be zero. This is usually satisfied by including an intercept in the regression model and ensures that there's no systematic bias in the residuals.

results and discusstion Correlation Coefficient (R) is large enough means results are reliable

R-squared (R^2) higher than 0.8 means results are reliable

Standard Error of Coefficients is small enough means results are reliable

p-value: If the p-value is below a specified threshold (e.g., 0.05), we can reject the null hypothesis, suggesting that the predictor variable significantly contributes to explaining the variation in the dependent variable.

residuals 95%of results within critical lines and they form normal distribution means model is good

[ ]: