**MATHEMATICS- STATISTICS**

**1. Descriptive statistics**

Descriptive statistics involve the use of numerical and graphical methods to summarize and describe the main features of a data set. These statistics provide a simple summary of the sample and the observations within it. The main measures of descriptive statistics include:

1. **Measures of Central Tendency:**
   * **Mean (Average):** The sum of all values divided by the number of values.
   * **Median:** The middle value of a dataset when it is ordered.
   * **Mode:** The value that occurs most frequently in a dataset.
2. **Measures of Dispersion (Variability):**
   * **Range:** The difference between the maximum and minimum values.
   * **Variance:** The average of the squared differences from the mean.
   * **Standard Deviation:** The square root of the variance; it measures the average distance of each data point from the mean.
3. **Measures of Shape:**
   * **Skewness:** Indicates the asymmetry or lack of symmetry in a distribution.
   * **Kurtosis:** Measures the "tailedness" or sharpness of a distribution's peak.
4. **Percentiles and Quartiles:**
   * **Percentiles:** Divide a dataset into 100 equal parts; the 50th percentile is the median.
   * **Quartiles:** Divide a dataset into four equal parts; the three quartiles are Q1, Q2 (median), and Q3.
5. **Graphical Representations:**
   * **Histograms:** Display the distribution of a continuous variable.
   * **Box Plots (Box-and-Whisker Plots):** Show the distribution of a dataset and highlight important summary statistics.
   * **Scatter Plots:** Display the relationship between two variables.

Descriptive statistics are essential for summarizing and understanding the main characteristics of a dataset before more complex analyses are performed. If you have specific questions or if there's a particular aspect of descriptive statistics you'd like more information on, feel free to ask!

**2. Bivariate analysis of factors that are significant to Catastrophic Expenditure at 5% (Catas5i)**

Bivariate analysis involves the examination of the relationship between two variables. In your case, you are interested in exploring factors that are significant to Catastrophic Expenditure at a 5% significance level (Catas5i). Here's a general outline of steps you might take:

1. **Identify Variables:**
   * Determine the variables you want to analyze in relation to Catastrophic Expenditure.
   * These variables could be demographic, economic, health-related, or any other factors that you suspect may be linked to Catastrophic Expenditure.
2. **Data Cleaning:**
   * Ensure that your data is clean and free of errors.
   * Handle missing data appropriately (impute or remove, depending on the extent of missingness).
3. **Descriptive Statistics:**
   * Calculate descriptive statistics for Catastrophic Expenditure and other relevant variables.
   * This can include means, medians, standard deviations, and other measures.
4. **Bivariate Analysis Techniques:**
   * **Correlation Analysis:** Use correlation coefficients (e.g., Pearson's correlation) to measure the strength and direction of linear relationships between continuous variables.
   * **T-Tests or ANOVA:** If your dependent variable is categorical and your independent variable is continuous, you might use t-tests or analysis of variance (ANOVA) to compare means across different groups.
   * **Chi-Square Test:** If both variables are categorical, you could use a chi-square test to examine the association between them.
5. **Visualization:**
   * Create visualizations such as scatter plots, box plots, or bar charts to visually explore relationships between variables.
6. **Hypothesis Testing:**
   * Formulate hypotheses regarding the relationship between the variables.
   * Conduct hypothesis tests to determine if the observed relationships are statistically significant.
   * Use a significance level of 5% (0.05) to assess statistical significance.
7. **Interpretation:**
   * Interpret the results in the context of your study.
   * Discuss the practical significance of the findings.
8. **Control for Confounding Variables:**
   * If relevant, consider controlling for confounding variables that might influence the relationship between your independent and dependent variables.
9. **Documentation:**
   * Clearly document your methods, results, and interpretations for transparency and replicability.

Remember that correlation does not imply causation, so be cautious in drawing causal conclusions from bivariate analyses alone. If you find significant associations, further multivariate analysis may be needed to control for other factors and strengthen the evidence for a causal relationship.

**3. Logistic regression analysis of predictors of Catastrophic Expenditure at 5% (Catas5i) Independent variables: q101 : SD/NSD q103: public/private facility q104: Secondary/tertiary facility q110\_pat\_recode: Age of patient recoded q111 Sex of patient q300Fpayer\_2gps Ethnicity 2gps q300Gpayer Religion Dependent Variable catas5i Catastrophic Expenditure at 5% threshold**

Logistic regression is a statistical method used for predicting the probability of an event occurring. In your case, you want to use logistic regression to analyze predictors of Catastrophic Expenditure at a 5% threshold (Catas5i). Here's a step-by-step guide on how to perform logistic regression using the variables you've listed:

**Steps:**

1. **Data Preparation:**
   * Ensure that your data is cleaned and properly formatted.
   * Recode categorical variables into dummy/indicator variables if necessary.
2. **Variable Definitions:**
   * Identify your independent variables (predictors) and your dependent variable (Catas5i).
3. **Dummy Coding:**
   * If your categorical variables are not already dummy-coded, create dummy variables for them. This is especially important for variables like facility type, patient age group, sex, ethnicity, religion, etc.
4. **Data Splitting:**
   * Split your dataset into a training set and a testing set. The training set will be used to build the model, and the testing set will be used to evaluate its performance.
5. **Logistic Regression Model:**
   * Fit a logistic regression model using software (e.g., Python with scikit-learn, R, or other statistical software).
   * In Python with scikit-learn, the code might look like this:

python

from sklearn.model\_selection import train\_test\_split from sklearn.linear\_model import Logistic Regression from sklearn.metrics import accuracy\_score, classification\_report

# Assuming X contains your independent variables and y is the dependent variable X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Create and fit the logistic regression model model = LogisticRegression() model.fit(X\_train, y\_train)

# Make predictions on the test set y\_pred = model.predict(X\_test)

# Evaluate the model accuracy = accuracy\_score(y\_test, y\_pred) print(f'Accuracy: {accuracy}') print(classification\_report(y\_test, y\_pred))

1. **Interpretation:**
   * Interpret the coefficients of the logistic regression model.
   * The coefficients tell you the direction and strength of the relationship between each independent variable and the log-odds of Catastrophic Expenditure.
2. **Significance Testing:**
   * Assess the significance of the coefficients and overall model fit.
   * Use a significance level of 5% to determine statistical significance.
3. **Model Evaluation:**
   * Assess the performance of your model using metrics like accuracy, precision, recall, and the ROC curve.
4. **Discussion:**
   * Discuss the results in the context of your research question.
   * Consider the practical significance of the findings.

Remember that logistic regression assumes a linear relationship between the log-odds of the dependent variable and the independent variables. If your data suggests a nonlinear relationship, you may need to explore more complex models or transformations of variables. Additionally, be cautious about overfitting the model to the training data; validation techniques can help address this concern.