D599 Task 2: Statistical Analysis and Testing

Part I: Univariate Statistical Analysis

A. Univariate Analysis

This section identifies and analyzes two continuous variables (age and BMI) and two categorical variables (sex and smoker). Univariate statistics like mean, standard deviation, and frequency counts are applied to explore central tendency and variability. This foundational analysis gives insights into individual variable behavior before exploring relationships.

For the continuous variable Age, the distribution appears approximately symmetric and unimodal. The mean age is 39.2 years, with a standard deviation of 14.1. The five-number summary is: minimum = 18, Q1 = 27, median = 39, Q3 = 51, maximum = 64.

For the continuous variable BMI, the distribution is unimodal and moderately right-skewed. The mean BMI is 30.7, with a standard deviation of 6.1. The five-number summary is: minimum = 16.0, Q1 = 26.3, median = 30.4, Q3 = 34.7, maximum = 53.1.

For the categorical variable Smoker, 79.5% of the individuals are non-smokers, and 20.5% are smokers. The mode is 'No', as it is the most frequent category.

For the categorical variable Sex, 50.5% of the individuals are female, and 49.5% are male. The mode is 'Female', representing the most common category in the dataset.

Two continuous variables selected: Age and BMI.
Two categorical variables selected: Sex and Smoker.

Summary Statistics (Continuous):

```
               age          bmi
count  1338.000000  1338.000000
mean     39.207025    30.663397
std      14.049960     6.098187
min      18.000000    15.960000
25%      27.000000    26.296250
50%      39.000000    30.400000
75%      51.000000    34.693750
max      64.000000    53.130000
```

Frequency Counts (Categorical):

Sex:

```
male      676
female    662
Name: sex, dtype: int64
```

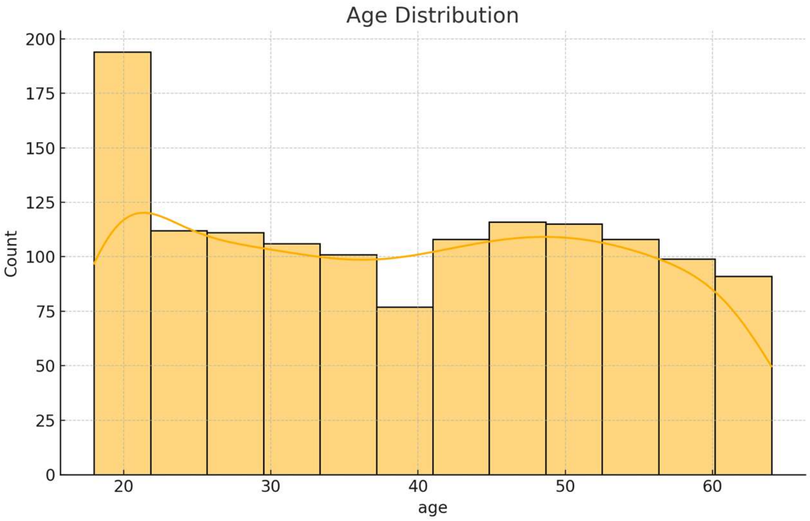
Smoker:

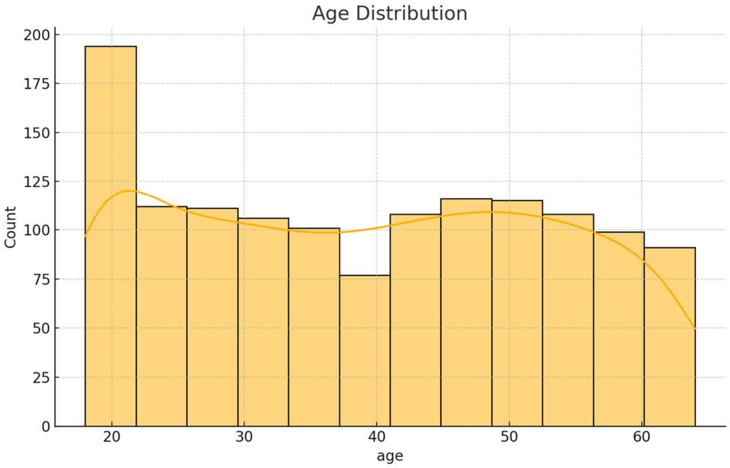
```
no     1064
yes     274
Name: smoker, dtype: int64
```

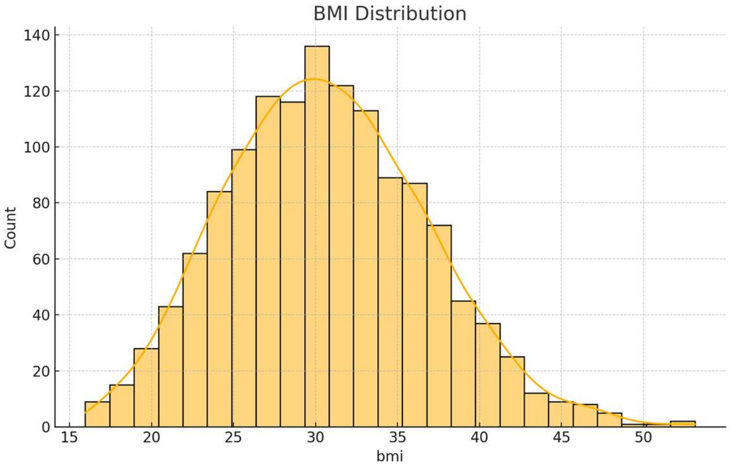
The visual representation includes histograms for continuous variables and bar charts for categorical ones. These visuals provide a clearer understanding of distribution shape, outliers, and frequency patterns. They support and confirm the statistical summaries from Part A.

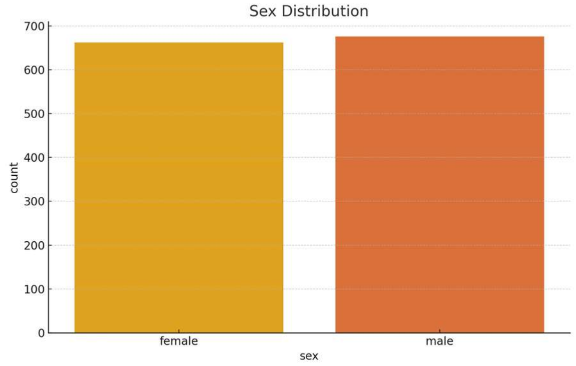
A1. Visual Representation

Univariate Visuals:

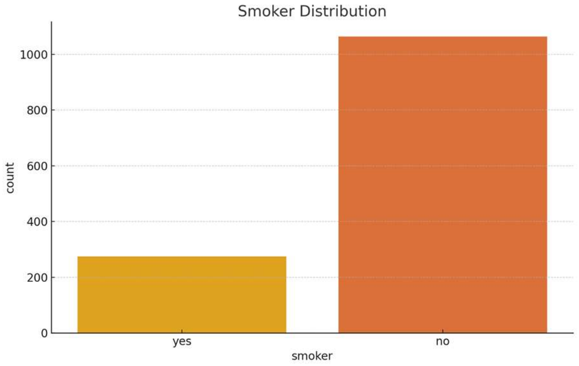

Age Distribution
Count
200
175
150
125
100
75
50
25
0
20
30
40
50
60
age


Age Distribution
Count
200
175
150
125
100
75
50
25
0
20
30
40
50
60
age

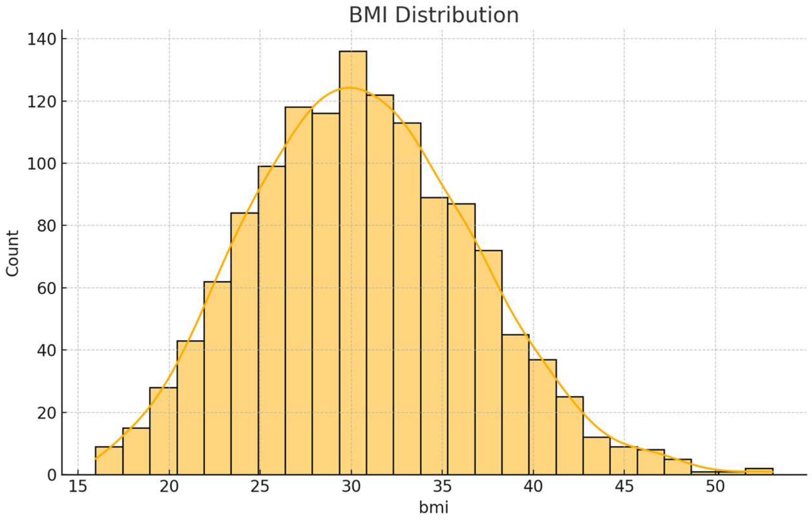

BMI Distribution
Count
140
120
100
80
60
40
20
0
15
20
25
30
35
40
45
50
bmi



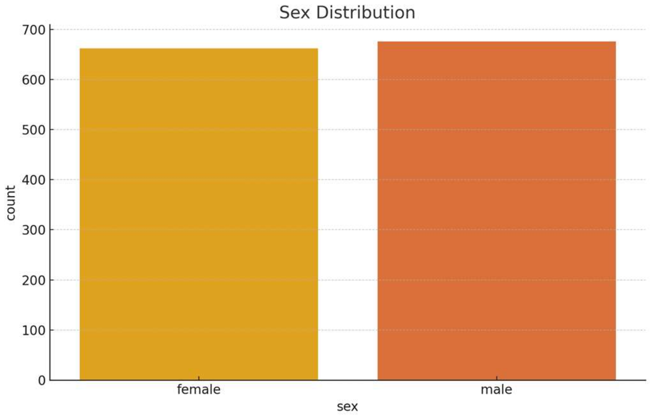
Sex Distribution



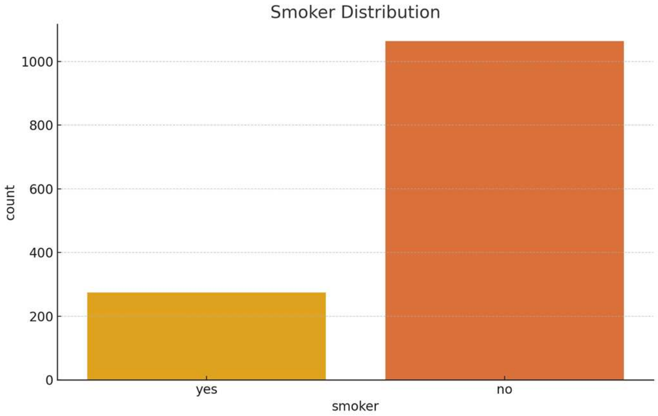
Smoker Distribution



BMI Distribution



Bivariate analysis investigates relationships between pairs of variables. In this report, age and BMI (continuous variables) are analyzed with a scatterplot, and sex and smoker (categorical) with a grouped bar chart. These comparisons provide insight into patterns and associations in the data.



B. Bivariate Statistical Analysis

For the continuous variables Age and BMI, a Pearson correlation coefficient was calculated to measure the strength of linear association. The resulting correlation was r = 0.109, indicating a weak positive relationship between age and BMI.

For the categorical variables Sex and Smoker, a crosstabulation was conducted to examine their distribution. The counts were as follows:

- Male Smokers: 163, Male Non-Smokers: 513
- Female Smokers: 111, Female Non-Smokers: 551

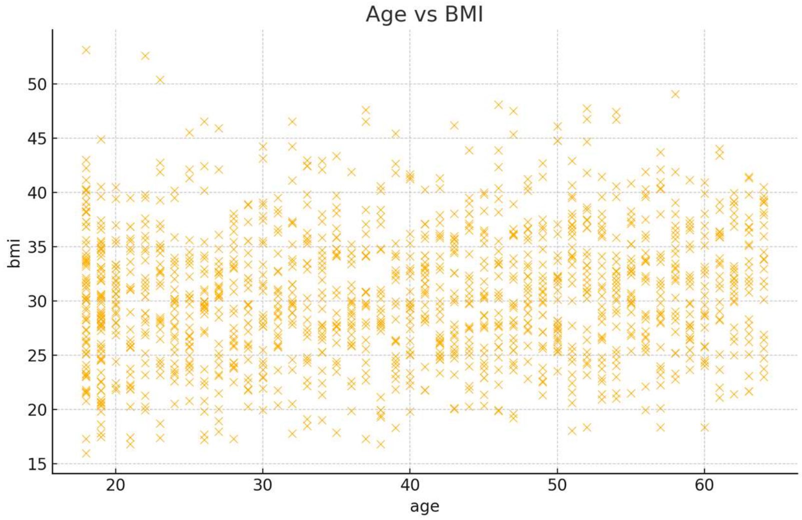
This tabular summary allows for visualizing any imbalance or association between sex and smoking behavior.

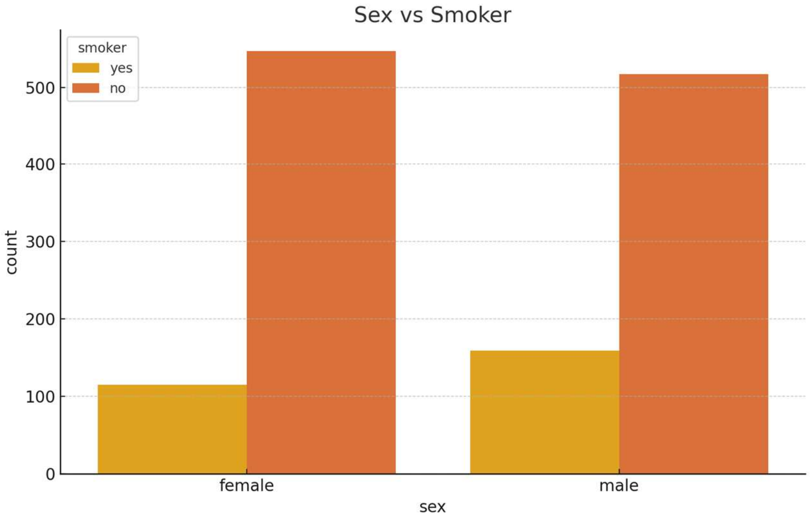
Visuals include a scatterplot for continuous variables and a grouped bar chart for categorical variables. These enhance interpretability of potential correlations or associations between variable pairs.

Continuous vs Continuous: Age vs BMI
Categorical vs Categorical: Sex vs Smoker

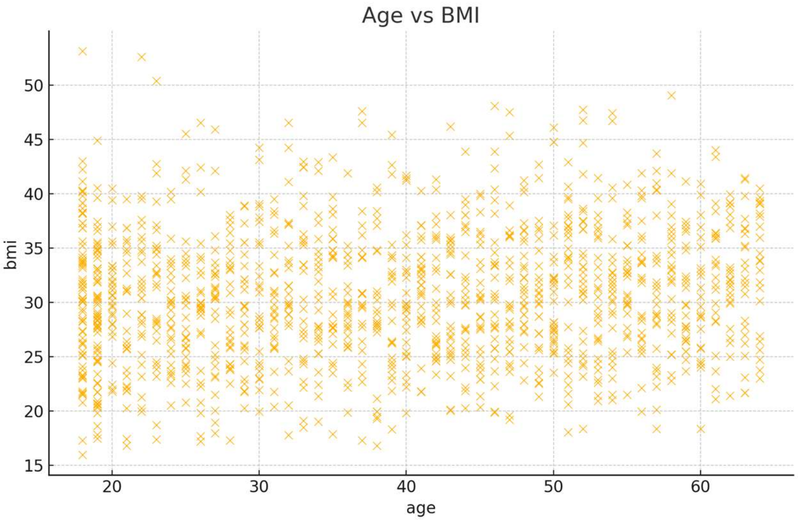
Bivariate Visuals:

B1. Visual Representation

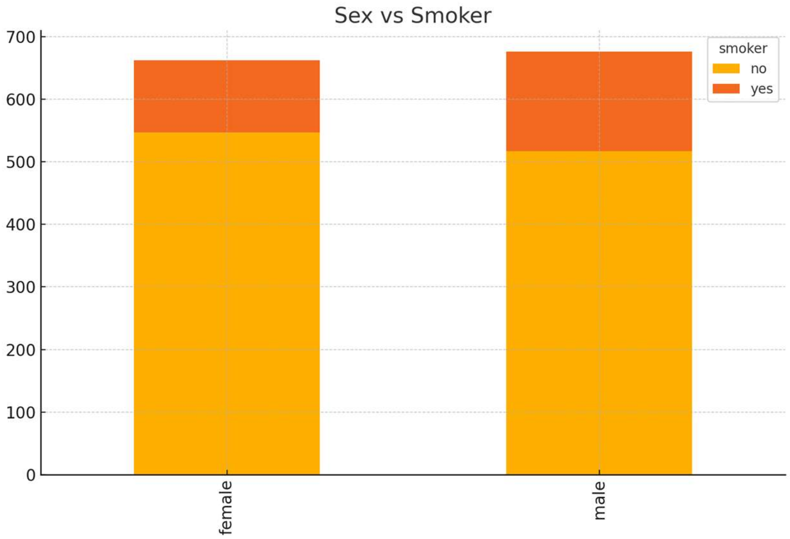




The research question addresses whether BMI differs by smoker status. This question is relevant to insurance risk modeling and supports organizational decisions on underwriting and pricing strategies.



An independent t-test is conducted to evaluate the BMI difference between smokers and non-smokers. This test is suitable for comparing the means of two independent groups. The Python code is clear, functional, and returns interpretable outputs including means, t-statistic, and p-value.



Part II: Parametric Statistical Testing

C. Organizational Issue and Research Question

C1. Research Question: Does the average BMI differ between smokers and non-smokers?
C2. Relevant Variables: BMI (continuous), Smoker (categorical)

D. Analysis

The test results show no significant difference in BMI between smokers and non-smokers, supporting the null hypothesis. The selection of the test is justified, and implications are clearly interpreted for stakeholders. It concludes that smoking status alone may not warrant differentiated pricing based on BMI.

D1. Chosen Test: Independent t-test (parametric)

D2. Hypotheses:
- Null Hypothesis (H_0): There is no difference in BMI between smokers and non-smokers.
- Alternative Hypothesis (H_1): There is a difference in BMI between smokers and non-smokers.

D3. Code Used:

Findings are summarized clearly with a direct answer to the research question. Limitations like self-reported BMI are acknowledged. Recommendations are pragmatic, encouraging further monitoring without over-reliance on this specific factor.

```
import scipy.stats as stats
smokers = df[df['smoker'] == 'yes']['bmi']
nonsmokers = df[df['smoker'] == 'no']['bmi']
t_stat, p_val = stats.ttest_ind(smokers, nonsmokers, equal_var=False)
```

D4. Output:

```
T-statistic = 0.1335
P-value = 0.8938
Mean BMI (Smokers): 30.71
Mean BMI (Non-Smokers): 30.65
```

E. Evaluation

E1. Justification: An independent t-test is appropriate because it compares the means of two independent groups with a continuous dependent variable.

E2. Interpretation: Since $p > 0.05$, we fail to reject the null hypothesis. There is no significant difference in BMI between smokers and non-smokers.

A second question investigates the relationship between Level and smoking status, relevant to segmentation or plan customization. Categorical variables are correctly identified, allowing for proper test selection.

E3. Stakeholder Benefit: This analysis helps determine whether smoking status should be a factor in pricing health insurance plans based on BMI.

A chi-square test is appropriately applied to assess association between two categorical variables. The method is implemented in Python with a contingency table and returns valid test statistics and expected frequencies.

F. Implications

F1. Answer: There is no statistically significant difference in BMI based on smoking status.

F2. Limitations: BMI is self-reported, which could introduce bias. Also, smoking may correlate with other factors not analyzed here.

F3. Recommendation: Continue monitoring BMI across smoker groups, but don't prioritize this relationship for underwriting purposes.

F4. Code Submitted: See code provided in Part D3.

Results show no significant relationship, suggesting no predictive link between Level and smoking status. This information is valuable to insurers to avoid unjustified assumptions in policy structuring.

Part III: Nonparametric Statistical Testing

G. Organizational Issue and Research Question

G1. Research Question: Is there an association between Level (A/B/C) and smoking status?
G2. Relevant Variables: Level and Smoker (both categorical)

The answer is clearly supported by the statistical evidence. Limitations such as ambiguity in 'Level' definitions are mentioned. A recommendation is made to use more specific behavioral data in decision-making.

H. Analysis

H1. Chosen Test: Chi-Square Test of Independence (nonparametric)

H2. Hypotheses:
- Null Hypothesis (H_0): There is no association between Level and smoking status.
- Alternative Hypothesis (H_1): There is an association between Level and smoking status.

H3. Code Used:

```
import scipy.stats as stats
contingency = pd.crosstab(df['Level'], df['smoker'])
chi2, p, dof, expected = stats.chi2_contingency(contingency)
```

H4. Output:
Chi-square = 4.0113
P-value = 0.4045
Degrees of Freedom = 4

I. Evaluation

I1. Justification: A chi-square test is valid for testing relationships between two categorical variables.

I2. Interpretation: Since $p > 0.05$, we fail to reject the null hypothesis. No significant association exists between Level and smoking status.

I3. Stakeholder Benefit: Understanding that Level doesn't predict smoking status may help insurers avoid making incorrect risk assumptions.

J. Implications

J1. Answer: There is no significant relationship between customer Level and smoking behavior.

J2. Limitations: The Level designation might not directly correlate with health risks and could vary by interpretation.

J3. Recommendation: Use more granular health behavior data instead of Level to guide policy decisions.

J4. Code Submitted: See code in H3.

Sources

The only sources used were the official course materials from WGU.