Open in Colab

Problem Set: Analyzing Test Scores Data

Chi-Squared Test: Race and Lunch

We will perform a chi-squared test to investigate the relationship between 'race' and 'lunch' in the dataset.

Null Hypothesis (H_0): There is no statistically significant relationship between race and lunch. (Race and lunch are independent).

Alternative Hypothesis (H_1): There is a statistically significant relationship between race and lunch. (Race and lunch are dependent).

Significance Level (α): 0.05

Statistical Statements:

- If the p-value is less than or equal to α ($p \leq 0.05$), we reject the null hypothesis and conclude that there is a statistically significant relationship between race and lunch.
- If the p-value is greater than α ($p > 0.05$), we fail to reject the null hypothesis and conclude that there is no statistically significant relationship between race and lunch.

In [ ]:

```
```

In [ ]:

```
import pandas as pd
import numpy as np
import scipy as sp
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import chi2_contingency
from google.colab import files
uploaded = files.upload()
for fn in uploaded.keys():
  print('User uploaded file "{name}" with length {length} bytes'.format(
      name=fn, length=len(uploaded[fn])))
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

In [ ]:

```
# Assuming the file uploaded was 'Wk 6 S 2 Test scores.xlsx'
df = pd.read_excel('Wk 6 S 2 Test scores.xlsx')

print("DataFrame shape:")
display(df.shape)
print("\nDataFrame head:")
```

```python
display(df.head())
print("\nDataFrame info:")
display(df.info())
print("\nDataFrame describe:")
display(df.describe())
```

```
DataFrame shape:
(250, 9)
DataFrame head:
```

	ID	read	math	class	experience	sex	lunch	race	schoolnum
0	1	445	475	small.class	9	girl	no	white	4
1	2	447	539	small.class	19	girl	no	black	2
2	3	440	465	regular.with.aide	0	boy	yes	black	1
3	4	447	557	regular	14	boy	no	white	4
4	5	445	490	small.class	6	boy	yes	white	4

```
DataFrame info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   ID          250 non-null    int64 
 1   read        250 non-null    int64 
 2   math        250 non-null    int64 
 3   class       250 non-null    object
 4   experience  250 non-null    int64 
 5   sex         250 non-null    object
 6   lunch       250 non-null    object
 7   race        250 non-null    object
 8   schoolnum   250 non-null    int64 
dtypes: int64(5), object(4)
memory usage: 17.7+ KB
None
DataFrame describe:
```

	ID	read	math	experience	schoolnum
count	250.000000	250.000000	250.000000	250.00000	250.000000
mean	125.500000	435.344000	489.204000	8.90000	2.416000
std	72.312977	29.283027	42.354907	5.80351	1.098988
min	1.000000	384.000000	401.000000	0.00000	1.000000
25%	63.250000	415.000000	460.000000	4.00000	1.250000
50%	125.500000	432.500000	483.500000	9.00000	2.000000
75%	187.750000	448.000000	515.750000	13.00000	3.000000
max	250.000000	605.000000	622.000000	27.00000	5.000000

In [ ]:

```
# Count 'yes' and 'no' values in the 'lunch' column
lunch_counts = df['lunch'].value_counts()
print("Counts of 'yes' and 'no' in 'lunch' column:")
display(lunch_counts)

# Group by 'race' and count 'yes' and 'no' in 'lunch' for each race
lunch_counts_by_race = df.groupby('race')['lunch'].value_counts().unstack(fill_valu
print("\nCounts of 'yes' and 'no' in 'lunch' column grouped by 'race':")
display(lunch_counts_by_race)
```

```
Counts of 'yes' and 'no' in 'lunch' column:
```

	count
lunch	
no	132
yes	118

dtype: int64

```
Counts of 'yes' and 'no' in 'lunch' column grouped by 'race':
```

lunch	**no**	**yes**
race		
black	28	64
white	104	54

In [ ]:

```
# Create a contingency table of 'race' and 'lunch'
contingency_table = pd.crosstab(df['race'], df['lunch'])
print("Contingency Table:")
display(contingency_table)

# Perform the chi-squared test
chi2, p, dof, expected = chi2_contingency(contingency_table)
```

```
print(f"\nChi-squared statistic: {chi2}")
print(f"P-value: {p}")
print(f"Degrees of freedom: {dof}")
print("Expected frequencies:")
display(pd.DataFrame(expected, index=contingency_table.index, columns=contingency_t

# Interpret the results
alpha = 0.05
print("\nInterpretation:")
if p <= alpha:
    print(f"Since the p-value ({p:.4f}) is less than or equal to the significance l
    print("Conclusion: There is a statistically significant relationship between ra
else:
    print(f"Since the p-value ({p:.4f}) is greater than the significance level ({al
    print("Conclusion: There is no statistically significant relationship between r
```

```
Contingency Table:
```

lunch	no	yes
race		
black	28	64
white	104	54

```
Chi-squared statistic: 27.814646654859555
P-value: 1.3351153144840636e-07
Degrees of freedom: 1
Expected frequencies:
```

lunch	no	yes
race		
black	48.576	43.424
white	83.424	74.576

```
Interpretation:
Since the p-value (0.0000) is less than or equal to the significance level (0.05), w
e reject the null hypothesis.
Conclusion: There is a statistically significant relationship between race and lunc
h.
```

In [ ]:

It appears that a higher proportion of students who recieve free lunches are Black compared to those who are white.

Chi-Squared Test: Math and Reading Scores

We will attempt to perform a chi-squared test to investigate the relationship between 'math' and 'read' scores in the dataset.

Note: Chi-squared tests are designed for categorical variables. To perform this test on numerical scores like math and reading, the scores would typically need to be converted into categories (e.g., by binning).

Null Hypothesis (H_0): There is no statistically significant relationship between math and reading scores. (Math and reading scores are independent).

Alternative Hypothesis (H_1): There is a statistically significant relationship between math and reading scores. (Math and reading scores are dependent).

Significance Level (α): 0.05

Statistical Statements:

- If the p-value is less than or equal to α ($p \leq 0.05$), we reject the null hypothesis and conclude that there is a statistically significant relationship between math and reading scores.
- If the p-value is greater than α ($p > 0.05$), we fail to reject the null hypothesis and conclude that there is no statistically significant relationship between math and reading scores.

In [10]:

```
# Bin 'math' and 'read' scores into 'low' and 'high' groups
# We can use the median as a simple split point for demonstration
math_median = df['math'].median()
read_median = df['read'].median()

df['math_category'] = pd.cut(df['math'], bins=[-float('inf'), math_median, float('i
df['read_category'] = pd.cut(df['read'], bins=[-float('inf'), read_median, float('i

print("Counts for Math Categories:")
display(df['math_category'].value_counts())

print("\nCounts for Reading Categories:")
display(df['read_category'].value_counts())

print("\nCross-tabulation of Math and Reading Categories:")
display(pd.crosstab(df['math_category'], df['read_category']))

# Display the first few rows with the new categories
print("\nDataFrame head with new categories:")
display(df.head())
```

```
Counts for Math Categories:
```

	count
math_category	
low	125
high	125

dtype: int64

```
Counts for Reading Categories:
```

	count
read_category	
low	125
high	125

dtype: int64

```
Cross-tabulation of Math and Reading Categories:
```

read_category	low	high
math_category		
low	91	34
high	34	91

```
DataFrame head with new categories:
```

	ID	read	math	class	experience	sex	lunch	race	schoolnum	math_catego
0	1	445	475	small.class	9	girl	no	white	4	lc
1	2	447	539	small.class	19	girl	no	black	2	hiç
2	3	440	465	regular.with.aide	0	boy	yes	black	1	lc
3	4	447	557	regular	14	boy	no	white	4	hiç
4	5	445	490	small.class	6	boy	yes	white	4	hiç

In [11]:
```python
from scipy.stats import chi2_contingency

# Create a contingency table of 'math_category' and 'read_category'
contingency_table_scores = pd.crosstab(df['math_category'], df['read_category'])
print("Contingency Table for Math and Reading Categories:")
display(contingency_table_scores)

# Perform the chi-squared test
chi2_scores, p_scores, dof_scores, expected_scores = chi2_contingency(contingency_t

print(f"\nChi-squared statistic for Math and Reading Categories: {chi2_scores}")
```

```
print(f"P-value for Math and Reading Categories: {p_scores}")
print(f"Degrees of freedom for Math and Reading Categories: {dof_scores}")
print("Expected frequencies for Math and Reading Categories:")
display(pd.DataFrame(expected_scores, index=contingency_table_scores.index, columns

# Interpret the results
alpha = 0.05
print("\nInterpretation for Math and Reading Categories:")
if p_scores <= alpha:
    print(f"Since the p-value ({p_scores:.4f}) is less than or equal to the signifi
    print("Conclusion: There is a statistically significant relationship between ma
else:
    print(f"Since the p-value ({p_scores:.4f}) is greater than the significance lev
    print("Conclusion: There is no statistically significant relationship between m
```

```
Contingency Table for Math and Reading Categories:
```

read_category	low	high
math_category		
low	91	34
high	34	91

```
Chi-squared statistic for Math and Reading Categories: 50.176
P-value for Math and Reading Categories: 1.4055640543520576e-12
Degrees of freedom for Math and Reading Categories: 1
Expected frequencies for Math and Reading Categories:
```

read_category	low	high
math_category		
low	62.5	62.5
high	62.5	62.5

```
Interpretation for Math and Reading Categories:
Since the p-value (0.0000) is less than or equal to the significance level (0.05), w
e reject the null hypothesis.
Conclusion: There is a statistically significant relationship between math and readi
ng scores (based on the defined categories).
```

Conclusions

Based on the statistical analysis conducted:

1. **Relationship between Race and Lunch:** The chi-squared test of independence between 'race' and 'lunch' yielded a p-value of $p = p: .4f$. Given a significance level of $\alpha = 0.05$, we observe that $p \leq \alpha$. Therefore, we reject the null hypothesis and conclude there is statistically significant evidence of a relationship between race and lunch in this dataset.

2. **Relationship between Math and Reading Scores (Categorized):** The chi-squared test of independence between the categorized 'math_category' and 'read_category' variables resulted in a p-value of $p = p_scores:.4f$. With a significance level of $\alpha = 0.05$, we observe that $p \leq \alpha$. Thus, we reject the null hypothesis and conclude there is statistically significant evidence of a relationship between the categorized math and reading scores in this dataset.