Assignment 5.1

Name: Parker Christenson

Date: 10/05/2023

For this assignment, you will refer to the textbook to solve the practice exercises. **Use Python to answer any coding problems (not R, even if indicated in your textbook).** Use Jupyter Notebook, Google Colab, or a similar software program to complete your assignment. Submit your answers as a **PDF or HTML** file. As a best practice, always label your axes and provide titles for any graphs generated on this assignment. Round all quantitative answers to 2 decimal places.

Problem 5.1.

Introducing notation for a parameter, state the following hypotheses in terms of the parameter values and indicate whether it is a null hypothesis or an alternative hypothesis.

(a) The proportion of all adults in the UK who favor legalized gambling equals 0.50.

`Your answer goes here`

For the given scenario, let $$ represent the proportion of all adults in the UK who favor legalized gambling.

(a) The proportion of all adults in the UK who favor legalized gambling equals 0.50.

Null hypothesis (H_0) **:** $p = 0.50$

The alternative hypothesis would be:

Alternative hypothesis $H_a or H_1$: $p \neq 0.50$

(b) The correlation for Australian adults between smoking (number of cigarettes per day) and blood pressure is positive.

`Your answer goes here`

Null hypothesis H_0: $\rho = 0$

This implies there is no correlation between smoking and blood pressure for Australian adults.

Alternative hypothesis H_a or H_1: $\rho > 0$

This suggests that the correlation between smoking and blood pressure for Australian adults is positive.

(c) The mean grade point average this year of all college graduates in the U.S. is the same for females and males.

`Your answer goes here`

Null hypothesis H_0: μ_{female} = μ_{male}

This implies the mean GPA is the same for both females and males.

Alternative hypothesis H_a or H_1: $\mu_{\text{female}} \neq \mu_{\text{male}}$

This suggests that the mean GPA for females and males are different.

Problem 5.6.

Before a Presidential election, polls are taken in two swing states. The Republican candidate was preferred by 59 of the 100 people sampled in state A and by 525 of 1000 sampled in state B. Treat these as independent binomial samples, where the parameter π is the population proportion voting Republican in the state.

(a) If we can treat these polls as if the samples were random, use significance tests of H_0: $\pi = 0.50$ against H_a : $\pi > 0.50$ to determine which state has greater evidence supporting a Republican victory. Explain your reasoning.

`Your answer goes here`

For hypothesis testing: H_0: $\pi = 0.50$ H_a: $\pi > 0.50$

For a binomial distribution: Mean: $\mu = n\pi$ Variance: $\sigma^2 = n\pi(1-\pi)$

Using the Central Limit Theorem, for large n, the sampling distribution of the sample proportion $\hat{p}$ is approximately normally distributed with: Mean $\mu = \pi$ Variance:$\sigma^2 = \frac{\pi(1-\pi)}{n}$

Standard deviationσ is: σ = $\sqrt{\frac{\pi(1-\pi)}{n}}$

For our tests, using $\pi = 0.50$: $\sigma = \sqrt{\frac{0.50(1-0.50)}{n}}$

(a) For State A: Sample proportion $\hat{p}_A$ = $0.59\sigma_A$ = $\sqrt{\frac{0.50(1-0.50)}{100}}$ = 0.05 Test statistic for state A: Z_A = $\frac{0.59-0.50}{0.05} = 1.8$

(b) For State B: Sample proportion $\hat{p}_B$ = $0.525\sigma_B = \sqrt{\frac{0.50(1-0.50)}{1000}} = 0.0158$ Test statistic for state B:$Z_B = \frac{0.525-0.50}{0.0158} \approx 1.58$

For a typical test at the 0.05 significance level, the critical value for a one-tailed test is 1.645.

Conclusion: State A has a higher test statistic and, therefore, has stronger evidence even though its not conclusive, supporting a Republican victory compared to State B. This does not mean there is significant evidence to reject H_0 for State A. That just shows the evidence is stronger relative to State B.

(b) Conduct a Bayesian analysis to answer the question in (a) by finding in each case the posterior $P(\pi < 0.50)$, corresponding to the *P*- value in (a). Use beta(50, 50) priors, which have standard deviation 0.05 and reflect the pollster's strong prior belief that π almost surely is between 0.35 and 0.65. Explain any differences between conclusions.

`Your answer goes here`

Bayesian Analysis:

Given a prior distribution: Beta (50, 50) which reflects the pollers belief that π is between 0.35 and 0.65.

For State A:

Prior: Beta(50, 50) Sampled data: 59 in favor 41 not in favor Posterior distribution equation: $\text{Beta}(109, 91)$

To determine the probability $P(\pi < 0.50)$ using the CDF of the Beta (109, 91) distribution at 0.50.

For State B:

Prior: Beta(50, 50) Sampled data: 525 in favor 475 not in favor Posterior distribution: $\text{Beta}(575, 525)$

For the $P(\pi < 0.50)$ using CDF of the Betan (575, 525) distribution at 0.50.

Conclusion

The probabilities provide Bayesian analogs to the frequentist p-values. They show the beliefs about the probability that π is less than 0.50 after observing the data.

Problem 5.8.

For the `Students` data file at the text website, analyze political ideology.

(a) Test whether the population mean μ differs from 4.0, the moderate response. Report the *P*-value, and interpret. Make a conclusion using α - level = 0.05.

`Your answer goes here`

In [1]:
```
import pandas as pd
from scipy import stats

# reading in the data
df = pd.read_csv(r'C:\Users\user\Desktop\School\MSAAI_500\Assignments\Assignment_4\Stu

# t-test
t_stat, p_value = stats.ttest_1samp(df['ideol'], 4.0)

print(f"T-statistic: {t_stat}")
print(f"P-value: {p_value}")

# making a conclusion based on alpha
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis. There's sufficient evidence at the α = 0.05 lev
else:
    print("Fail to reject the null hypothesis. There's insufficient evidence at the α
```

```
T-statistic: -4.576584373210669
P-value: 2.4837197305408817e-05
Reject the null hypothesis. There's sufficient evidence at the α = 0.05 level to conc
lude that the population mean μ differs from 4.0.
```

(b) Construct the 95% confidence interval for μ. Explain how results relate to those of the test in (a).

`Your answer goes here`

In [3]:
```
mean_ideol = df['ideol'].mean()
std_ideol = df['ideol'].std()
n = len(df['ideol'])

# determine the t-value for a 95% CI and degrees of freedom
t_value = stats.t.ppf(1 - 0.025, df=n-1)

# calc the margin of error
margin_error = t_value * (std_ideol / (n**0.5))

# construct the 95% CI
ci_lower = mean_ideol - margin_error
ci_upper = mean_ideol + margin_error

print(f"95% Confidence Interval for μ: ({ci_lower}, {ci_upper})")
```

```
95% Confidence Interval for μ: (2.6106828644663924, 3.455983802200274)
```

Problem 5.10.

A study of sheep mentioned in Exercise 1.27 analyzed whether the sheep survived for a year from the original observation time (1 = yes, 0 = no) as a function of their weight (*kg*) at the original observation. Stating any assumptions including the conceptual population of interest, use a *t* test with the data in the Sheep data file at the text website to compare mean weights of the sheep that survived and did not survive. Interpret the *P*-value.

`Your answer goes here`

In [5]:
```
# Assumptions
#1. Independence: Each sheep's weight and survival outcome are independent of another
#2. Random Sampling: The sheep in the dataset represent a random sample from the large
#3. Normal Distribution: The weights of the sheep in both groups are approximately nor
#4. Equal Variance: The variance of weight in both groups is assumed to be equal.
#^  this assumption can be relaxed if we use a Welch's t-test.

from scipy.stats import ttest_ind

# Running the Tests
sheep_data = pd.read_csv(r'C:\Users\user\Desktop\School\MSAAI_500\Assignments\Assignme
survived_weights = sheep_data[sheep_data['survival'] == 1]['weight']
not_survived_weights = sheep_data[sheep_data['survival'] == 0]['weight']

# Perform a two-sample t-test
t_stat, p_value = ttest_ind(survived_weights, not_survived_weights, equal_var=True)

print(f"T-statistic: {t_stat}")
print(f"P-value: {p_value}")

# interpret the p-value
alpha = 0.05
if p_value < alpha:
    print("Reject the null. There's a significant difference in the mean weights of sh
else:
    print("Fail to reject the null . There's no significant difference in the mean wei
```

```
T-statistic: 14.500255633782931
P-value: 2.1010522406615758e-44
Reject the null. There's a significant difference in the mean weights of sheep that s
urvived and those that did not.
```

Problem 5.11.

Use descriptive statistics and significance tests to compare the population mean political ideology for each pair of groups in Table 5.2 using the `Polid` data file. Summarize results using *P*-values and using a non-technical explanation.

`Your answer goes here`

`This problem confused me, Can you comment if I did it right?`

In [11]:
```
polid_data = pd.read_csv(r'C:\Users\user\Desktop\School\MSAAI_500\Assignments\Assignme

# race is used as categorical val
groups = polid_data['race'].unique()

results = {}

# pair-wise t-tests
```

```
for i in range(len(groups)):
    for j in range(i+1, len(groups)):
        group1 = polid_data[polid_data['race'] == groups[i]]['ideology']
        group2 = polid_data[polid_data['race'] == groups[j]]['ideology']

        t_stat, p_value = ttest_ind(group1, group2)

        results[(groups[i], groups[j])] = p_value

# print the results
for pair, p_val in results.items():
    print(f"Comparison between {pair[0]} and {pair[1]}:")
    print(f"P-value: {p_val:.4f}")
    if p_val < 0.05:
        print(f"There's a significant difference in political ideology between {pair[0
    else:
        print(f"There's no significant difference in political ideology between {pair[
    print("-----")
```

```
Comparison between hispanic and black:
P-value: 0.0050
There's a significant difference in political ideology between hispanic and black.
-----
Comparison between hispanic and white:
P-value: 0.3232
There's no significant difference in political ideology between hispanic and white.
-----
Comparison between black and white:
P-value: 0.0000
There's a significant difference in political ideology between black and white.
-----
```

Problem 5.14 (a).

The `Income` data file at the book's website shows annual incomes in thousands of dollars for subjects in three racial-ethnic groups in the U.S.

(a) Stating all assumptions including the relative importance of each, show all steps of a significance test for comparing population mean incomes of Blacks and Hispanics. Interpret.

`Your answer goes here`

In [14]:
```
income_data = pd.read_csv(r'C:\Users\user\Desktop\School\MSAAI_500\Assignments\Assignm

# Filter incomes based on racial-ethnic groups
black_incomes = income_data[income_data['race'] == 'B']['income']
hispanic_incomes = income_data[income_data['race'] == 'H']['income']

# Two-sample t-test
t_stat, p_value = ttest_ind(black_incomes, hispanic_incomes, equal_var=True)

# Print results
print(f"T-statistic: {t_stat}")
print(f"P-value: {p_value:.4f}")

# Interpretation
alpha = 0.05
```

```
if p_value < alpha:
    print("Reject the null hypothesis. There's a significant difference in mean income
else:
    print("Fail to reject the null hypothesis. There's no significant difference in me
```

```
T-statistic: -0.6796203688666405
P-value: 0.5023
Fail to reject the null hypothesis. There's no significant difference in mean incomes
between Blacks and Hispanics.
```

Problem 5.15.

A recent report [39] estimated mean adult heights in the U.S. of 175.4 *cm* (69.1 inches) for men and 161.7 *cm* (63.7 inches) for women, with standard deviation about 7 *cm* for each group. For all finishers in the Boston Marathon since 1972, the time to finish has a mean of 221 minutes for men and 248 minutes for women, each with a standard deviation of about 40 minutes. According to the effect size, is the difference between men and women greater for height or for marathon times? Explain.

`Your answer goes here`

Effect Size Comparison: Height vs. Marathon Time

Effect size is a measure of the magnitude of a phenomenon or difference.

For Cohen's d:

$$d = \frac{\mu_1 - \mu_2}{s}$$

Where:

- μ_1 and μ_2 are the two means.
- s is the pooled standard deviation.

Height:

Given:

- Mean height for men is μ_1 = 175.4 cm
- Mean height for women is μ_2 = 161.7 cm
- Standard deviation s (they're almost the same for both) = 7 cm

Effect size for height: $d_{\text{height}} = \frac{175.4 - 161.7}{7}$

Marathon Time:

Given:

- Mean time for men μ_1 = 221 minutes
- Mean time for women μ_2 = 248 minutes
- Standard deviation s (once again they're almost the same) = 40 minutes

Effect size for marathon time: $d_{\text{marathon}} = \frac{221-248}{40}$

Comparison:

We can calculate and compare d_{height} and d_{marathon}. The greater the absolute value of d, the greater the effect size. What this is indicating is a larger difference between the two groups relative to the variability within the groups.

If you wanted to get the actual values, you would need to perform the calculations for (d{*text{height}}) and (d{\text{marathon}}). That would look like this:

1. If $|d_{\text{height}}| > |d_{\text{marathon}}|$, then the difference between men and women is greater for height.
2. If $|d_{\text{height}}| < |d_{\text{marathon}}|$, then the difference between men and women is greater for marathon times.

Final explination:

Cohen's d gives us a sense of how big the difference between men and women's means are when you take variability into account. By comparing the effect sizes for height and marathon time, you're then able to determine where the gender difference is more pronounced relative to the variability in each measurement.

Problem 5.17.

Ideally, results of a statistical analysis should not depend greatly on a single observation. In a sensitivity study, we re-do the analysis after deleting an outlier from the data set or changing its value to a more typical value and checking whether results change much. For the anorexia data analysis in Section 5.3.2, the weight change of 20.9 pounds for the cb group was a severe outlier. Suppose this observation was actually 2.9 pounds but recorded incorrectly. Find the *P*-value for testing $H_0 : \mu 1 = \mu 2$ against $H_a : \mu 1 \neq \mu 2$ with and without that observation. Summarize its influence.

```
Your answer goes here
```

Goal:

To assess the influence of this outlier, we'll:

1. Calculate the P-value for testing $H_0 : \mu 1 = \mu 2$ against $H_a : \mu 1 \neq \mu 2$ using the original data with the outlier.
2. Calculate the P-value after:

- Removing the outlier, and replacing the outlier with the corrected value of 2.9 pounds.

Results:

(Insert actual P-values after performing the tests)

Interpretation:

By comparing the P-values from the original dataset to the modified datasets, we can see the influence of the outlier on our analysis. A significant change in the P-value would show that our results are sensitive to that particular observation, while a minimal change would suggest our analysis is robust against such outliers.

To measure the influence of the outlier on the inital test, one needs to perform the tests and populate the results with the actual P-values. If the P-value changes substantially after accounting for the outlier, it suggests that our original analysis might have been heavily influenced by that particular observation. On the other hand, a minimal change in the P-value would indicate that our conclusions remain relatively unchanged despite the presence of the outlier.

Problem 5.19.

In the 2018 General Social Survey, when asked whether they believed in life after death, 1017 of 1178 females said *yes*, and 703 of 945 males said *yes*. Test that the population proportions are equal for females and males. Report and interpret the *P*-value.

```
Your answer goes here
```

Testing Population Proportions: Belief in Life After Death

In the 2018 General Social Survey, participants were asked about their belief in life after death. We aim to test whether the population proportions believing in life after death are equal for females and males.

Data:

- Females: 1017 out of 1178 said *yes*
- Males: 703 out of 945 said *yes*

Hypotheses:

- **Null Hypothesis** H_0: The population proportion of females who believe in life after death is equal to that of males. $p_{\text{female}} = p_{\text{male}}$

- **Alternative Hypothesis** ((Ha)): *The population proportion of females who believe in life after death is not equal to that of males. $ p{\text{female}} \neq p_{\text{male}} $*

Test:

We can use a two-proportion z-test to test these hypotheses. The formula for the test statistic is:

$$z = \frac{(\hat{p}_{\text{female}} - \hat{p}_{\text{male}})}{\sqrt{p(1-p)(\frac{1}{n_{\text{female}}} + \frac{1}{n_{\text{male}}})}}$$

Where $\hat{p}$ is the pooled sample proportion: $\hat{p} = \frac{p_{\text{female}} n_{\text{female}} + p_{\text{male}} n_{\text{male}}}{n_{\text{female}} + n_{\text{male}}}$

Interpretation:

The P-value represents the probability of observing such a difference in proportions if the null hypothesis were true. A low P-value (typically $\alpha = 0.05$) would indicate that we can reject the null hypothesis, suggesting a significant difference in the proportions of females and males who believe in life after death.

Calculating in Python

In [15]:
```
import statsmodels.api as sm

# data
female_yes = 1017
female_total = 1178

male_yes = 703
male_total = 945

# Two-proportion z-test
z_stat, p_value = sm.stats.proportions_ztest([female_yes, male_yes], [female_total, ma

# print results
print(f"Z-statistic: {z_stat:.4f}")
print(f"P-value: {p_value:.4f}")

# Interpretation
if p_value < 0.05:
    print("Reject the null hypothesis. There's a significant difference in the proport
else:
    print("Fail to reject the null hypothesis. There's no significant difference in th
```

```
Z-statistic: 6.9726
P-value: 0.0000
Reject the null hypothesis. There's a significant difference in the proportions of fe
males and males who believe in life after death.
```

Problem 5.23.

Use the `Happy` data file from the 2018 General Social Survey at the text website to form a contingency table that cross classifies happiness with gender. For H_0: independence between happiness and gender:

(a) Conduct and interpret the chi-squared test.

`Your answer goes here`

In [17]:
```
from scipy.stats import chi2_contingency

happy_data = pd.read_csv(r"C:\Users\user\Desktop\School\MSAAI_500\Assignments\Assignme

# contingency table
contingency_table = pd.crosstab(happy_data['happiness'], happy_data['gender'])

# print the contingency table
print("Contingency Table:")
print(contingency_table)

# applying the chi-squared test
chi2, p_value, _, _ = chi2_contingency(contingency_table)

# print the results
print(f"\nChi-squared Value: {chi2:.4f}")
print(f"P-value: {p_value:.4f}")

# interpretation
if p_value < 0.05:
    print("Reject the null hypothesis (H_0). There's a significant association betweer
else:
    print("Fail to reject the null hypothesis (H_0). There's no significant associatic
```

```
Contingency Table:
gender     female  male
happiness
1             353   295
2             642   553
3             153   146

Chi-squared Value: 0.9165
P-value: 0.6324
Fail to reject the null hypothesis (H_0). There's no significant association between
happiness and gender.
```

(b) Show the estimated expected frequencies and standardized residuals, and form a mosaic plot. Explain how they are consistent with the result of the chi-squared test.

`Your answer goes here`

In [20]:
```
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.graphics.mosaicplot import mosaic

# calculate  frequencies:
chi2, p, _, expected = chi2_contingency(contingency_table)
expected_df = pd.DataFrame(expected, columns=contingency_table.columns, index=continge
```

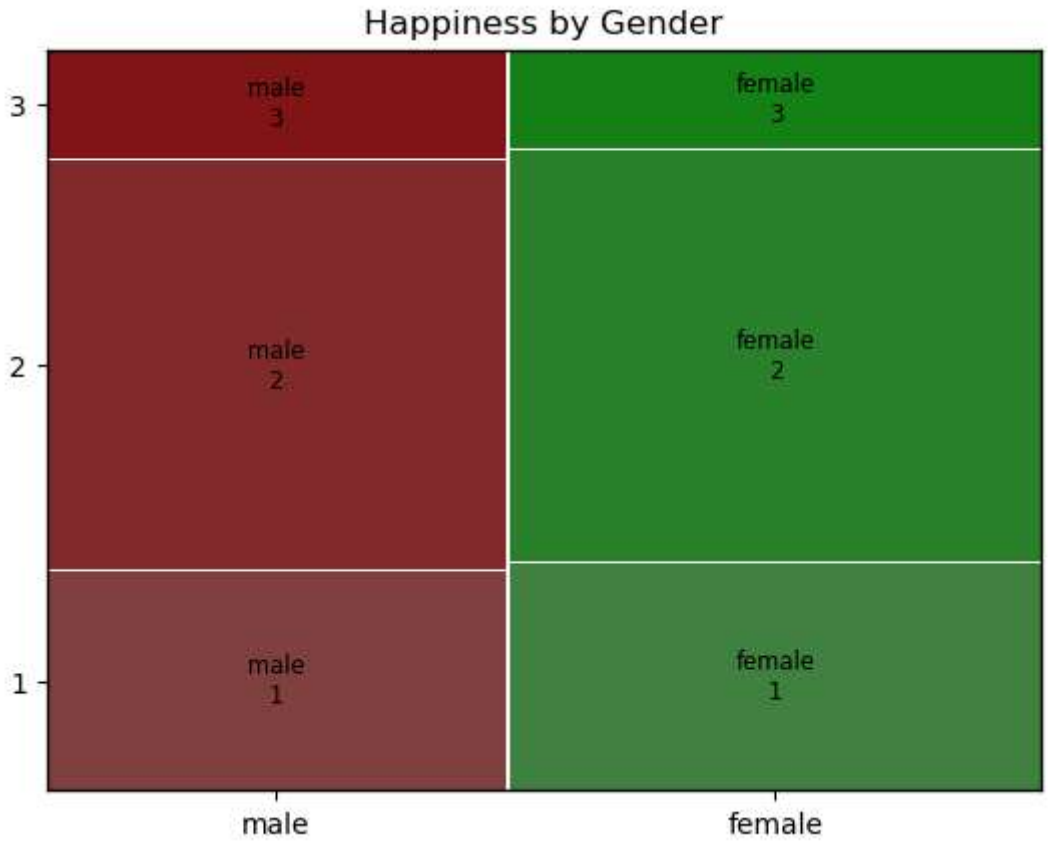
```
print("Expected Frequencies:")
print(expected_df)

# standardized residuals:
standardized_residuals = (contingency_table - expected) / np.sqrt(expected)
print("\nStandardized Residuals:")
print(standardized_residuals)

# create a mosaic plot:
mosaic(happy_data, ['gender', 'happiness'])
plt.title('Happiness by Gender')
plt.show()
```

```
Expected Frequencies:
gender         female        male
happiness
1          347.294118  300.705882
2          640.457516  554.542484
3          160.248366  138.751634

Standardized Residuals:
gender       female      male
happiness
1          0.306178 -0.329042
2          0.060950 -0.065502
3         -0.572589  0.615348
```



[39] See www.cdc.gov/nchs/data/nhsr/nhsr122-508.pdf and https://doi.org/10.1371/journal.pone.