Problem set 2

Due date and time: October 19, 23:30

(Please remember to include details of your analysis. If you use R or another programming language to perform the analysis or check your answer, please provide the code and screenshot displaying the results.)

1. (25 points) Fires are a common and important part of many ecosystems. Many species have evolved mechanisms for dealing with fire. Reed frogs, a species living in West Africa, have been observed hopping away from grass fires long before the heat of the fire reached the area they were in. This finding led to the hypothesis that the frogs might hear the fire and respond well before the fire reaches them. To test this hypothesis, researchers randomly selected 60 reed frogs and divided them into three groups. They then exposed each group to different sounds and recorded the frogs' responses (Grafe et al., 2002). Twenty frogs were exposed to the sound of fire, 20 to the sound of fire played backward (to control for the range of sound frequencies present in the real sound), and 20 to equally loud white noise. Of these 60 frogs, 17 hopped away from the sound of fire, 6 hopped away from the sound of fire played backward, and 2 hopped away from the white noise. Do the data provide evidence that reed frogs change their behavior in response to the sound of fire? Can a causal inference be made in this study, and if so, why or why not? (Please remember to state your null and alternative hypothesis, calculate test statistic and degree of freedom, determine if P-value is smaller than 0.05, and draw your conclusion.)

60 reed frogs
- 20 sound of the fire → 17
- 20 backward sound of the fire → 6
- 20 loud white noise → 2

	hopped away	No hopped
Sound of the fire	17	3
backward sound	6	14
White noise	2	18
Total	25	35

Chi-squared test (χ^2 contingency)
(Determine the difference between observed and expected data)

Null hypothesis (Ho):

The change of Reed frogs' behavior is not related to the sound of fire.

Alternative hypothesis (Ha):

The type of sound affects Reed frogs' behavior.

hw1.py × Release Notes: 1.94.2

hw1.py > ...

```python
import numpy as np
from scipy.stats import chi2_contingency

observed = np.array([[17, 3], [6, 14], [2, 18]])

chi2, p, dof, expected = chi2_contingency(observed)

print(f"Chi-squared stats : {chi2:.2f}")
print(f"P-value: {p:.5f}")
print(f"Degrees of freedom : {dof}")
print("Expected frequencies:")
print(expected)
```

Analysis (Assisted by ChatGPT)

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
 [ 8.33333333 11.66666667]]
(base) annieteng@sslvpn-252-075 hw_statistics % /Users/annieteng/miniforg
1.py
Chi-squared stats : 24.82
P-value: 0.00000
Degrees of freedom : 2
Expected frequencies:
[[ 8.33333333 11.66666667]
 [ 8.33333333 11.66666667]
 [ 8.33333333 11.66666667]]
(base) annieteng@sslvpn-252-075 hw_statistics %
```

Results:

Dof : 2

→ $df = (\text{row}-1)(\text{column}-1) = 2$ (3 rows, 2 columns)

Chi-squared statistics: 24.82

(The difference between observed and expected values is large)

P-value ≈ 0 (< 0.05)

→ Reject Ho

(There is a strong evidence to suggest the reed frogs change their behavior in reponse to the sound of fire.)

However, causal inference cannot be made in this study because it is an observational study not experimental. More variables are meant to be considered.

2. (25 points)Postnatal depression affects approximately 8–15% of new mothers. One theory about the onset of postnatal depression predicts that it may result from the stress of a complicated delivery. If so, then the rates of postnatal depression could be affected by the type of delivery. A study (Patel et al. 2005) of 10,934 women compared the rates of postnatal depression in mothers who delivered vaginally to those who had voluntary cesarean sections (C-sections). Of the 10,545 women who delivered vaginally, 1025 suffered significant postnatal depression. Of the 389 who delivered by voluntary C-section, 48 developed postnatal depression.

   (a) How different are the odds of depression under the two procedures? Calculate the odds ratio of developing depression, comparing vaginal birth to C-section.

   (b) Calculate a 95% confidence interval for the odds ratio.

   (c) Based on your result in part (b), would the null hypothesis that postpartum depression is independent of the type of delivery likely be rejected if tested?

   (d) What is the relative risk of postpartum depression under the two procedures? Compare your estimate to the odds ratio calculated in part (a).

   (e) Can a causal inference be made in this study, and if so, why or why not?

	Vaginally	C-section
depression	a 1025	c 48
No depression	b 9520	d 341
	10545	389

(a). Odds Ratio

Odds of depression (Vaginally): $\frac{1025}{9520} = 0.108$

= (C-section): $\frac{48}{341} = 0.141$

$OR = \frac{0.108}{0.141} = 0.766$

(b). Confidence Interval

CI: $e^{\log(OR) \pm 1.96 \times SE(\log(OR))}$

$SE(\log(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} = \sqrt{\frac{1}{1025} + \frac{1}{9520} + \frac{1}{48} + \frac{1}{341}} = 0.158$

95 CI = [0.561, 1.042]

(c). OR is in 95CI

95CI for OR is close to 1: The type of delievery has little effect on postnatal depression.

(d). Relative Risk

$RR = \frac{\frac{1025}{10545}}{\frac{48}{389}} = \frac{0.097}{0.123} = 0.788$ (Slightly larger than OR)

(e). Causal inference

No, because this is an observational study not an experimental study. More variables are required to control or consider.

3. (25 points) The following data are from a laboratory experiment by Smallwood et al. (1998) in which liver preparations from five rats were used to measure the relationship between the administered concentration of taurocholate (a salt normally occurring in liver bile) and the unbound fraction of taurocholate in the liver.

Rat	Concentration (μM)	Unbound fraction
1	3	0.63
2	6	0.44
3	12	0.31
4	24	0.19
5	48	0.13

(a) (5 points) Calculate the correlation coefficient r between the taurocholate unbound fraction and the concentration.

(b) (5 points) Plot the relationship between the two variables in a graph. Is the bivariate normality assumption underlying the correlation coefficient met?

(c) (8 points) Examine the plot in part (b). The relationship appears to be maximally strong, yet the correlation coefficient you calculated in part (a) is not near the maximum possible value. Why not?

(d) (7 points) What steps would you take with these data to meet the assumptions of correlation analysis?

(a).

hw3.py

Users > annieteng > Desktop > hw_statistics > hw3.py > ...

```python
import numpy as np

concentration = np.array([3, 6, 12, 24, 48])
unbound_fraction = np.array([0.63, 0.44, 0.31, 0.19, 0.13])

r = np.corrcoef(concentration, unbound_fraction)[0, 1]

print(f"Correlation coefficient : {r}")
```

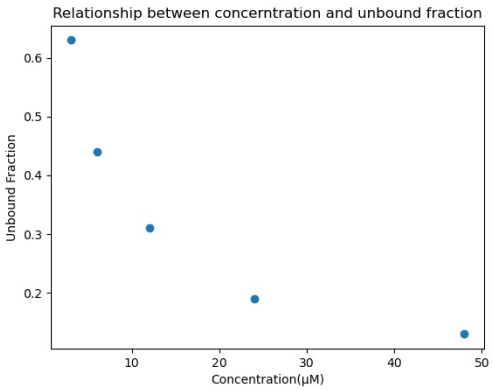
(Assisted by ChatGPT)

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS Python

```
(base) annieteng@sslvpn-252-037 ~ % /Users/annieteng/miniforge3/bin/python /Users/annieteng/Desktop/hw_statistics/hw3.py
Correlation coefficient : -0.8544602800107467
(base) annieteng@sslvpn-252-037 ~ %
```

$r = -0.854$

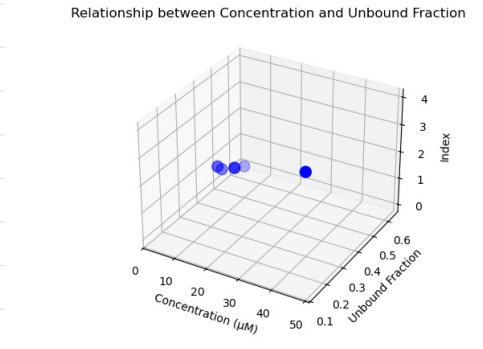
(b).



hw3.py

Users > annieteng > Desktop > hw_statistics > hw3.py > ...

```python
import numpy as np
import matplotlib.pyplot as plt

concentration = np.array([3, 6, 12, 24, 48])
unbound_fraction = np.array([0.63, 0.44, 0.31, 0.19, 0.13])

r = np.corrcoef(concentration, unbound_fraction)[0, 1]

print(f"Correlation coefficient : {r}")

plt.scatter(concentration, unbound_fraction)
plt.title('Relationship between concerntration and unbound fraction')
plt.xlabel('Concentration(μM)')
plt.ylabel('Unbound Fraction')

plt.show()
```



hw3.py

Users > annieteng > Desktop > hw_statistics > hw3.py > ...

```python
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

concentration = np.array([3, 6, 12, 24, 48])
unbound_fraction = np.array([0.63, 0.44, 0.31, 0.19, 0.13])

r = np.corrcoef(concentration, unbound_fraction)[0, 1]

x = concentration
y = unbound_fraction
z = np.arange(len(x))

fig, ax = plt.subplots(subplot_kw={"projection":"3d"})

ax.scatter(x, y, z, color='b', s=100)

ax.set_title('Relationship between Concentration and Unbound Fraction')
ax.set_xlabel('Concentration (μM)')
ax.set_ylabel('Unbound Fraction')
ax.set_zlabel('Index')

plt.show()
```

} Assisted by ChatGPT

From the 2D scatter plot and correlation coefficient r, we can observe that the relationship between concentration and unbound fraction is negative linear relationship.

But for the 3D plot, it's difficult to say that the scatter plot of X and Y has a circular or elliptical shape since the data points are insufficient.

⇒ I think it needs more data points or a further analysis to verify.

(c).

For a maximum correlation coefficient value: $\begin{cases} r = 1 \\ r = -1 \end{cases}$

It means that all data points should lie on a straight line and be perfectly linear. However, the points from the ⋈ scatter plot show some deviation from a straight line, which lowers the r from maximum value (-1).

⇒ It's may due to insufficient data points or some level of noise / natural variability

(d.)

Firstly, I will check for the linearity by using residual plot and find the outliers. However, the sample size is so small that I'm not sure about the necessity of removing the outliers. Anyway, if it violates normal distribution and linearity, then I will transform the data into logistic form. The overall purpose is to meet linearity and normal distribution.

4. (25 points) Large males of the European earwig, *Forficula auricularia*, develop abdominal forceps, which are used in fighting and courtship. Smaller males do not develop the forceps. Tomkins and Brown (2004) compared the proportion of males having forceps on islands in the North Sea with the population density of earwigs, measured as number caught per trap. Their data are listed in the following table.

Islands	Earwig density (number per trap)	Proportion of males with forceps
1	0.3	0.04
2	5.2	0.02
3	2.5	0.07
4	25.6	0.06
5	8.1	0.13
6	0.3	0.15
7	4.7	0.20
8	3.3	0.24
9	7.8	0.25
10	20.0	0.19
11	31.0	0.22
12	25.0	0.32
13	43.3	0.30
14	33.9	0.44
15	33.9	0.52
16	32.7	0.55
17	33.8	0.62
18	12.7	0.66
19	57.0	0.46
20	52.5	0.38
21	64.0	0.38
22	70.4	0.46

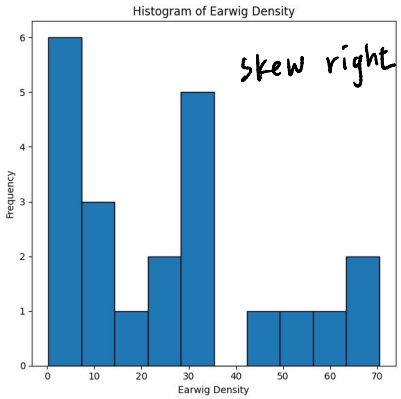
(a) (5 points) Plot the relationship between the two variables in a graph and explain why the distribution of the two variables is not bivariate normal. Try transforming the data (please provide at least one figure) and see if the transformation helps.

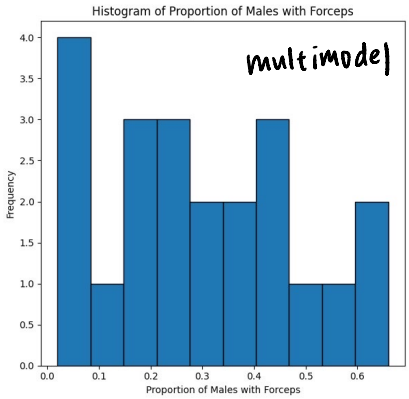
(b) (10 points) Choose the most appropriate method to test whether the two variables are correlated. Please remember to state your null and alternative hypothesis, calculate test statistic, determine if P-value is smaller than 0.05, and draw your conclusion. (Hint: The critical value corresponding to this test is 0.425. You can determine whether P-value is greater or smaller than 0.05 by comparing your test statistic to the critical value).

(c) (5 points) What are your assumptions in part (b)?

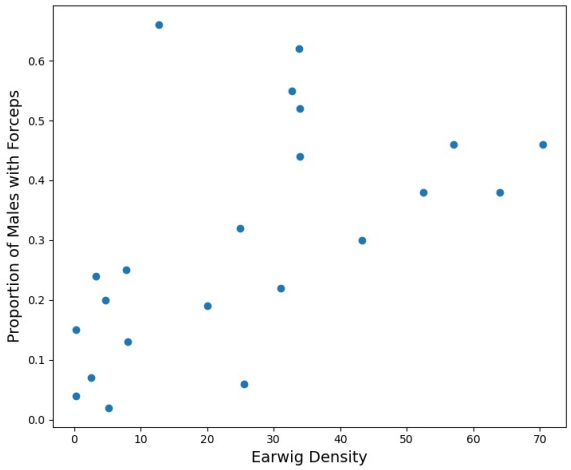
(d) (5 points) Earwig density on an island and the proportion of males with forceps are estimates, so the measurements of both variables include sampling error. In light of this fact, would the true correlation between the two variables tend to be larger, smaller, or the same as the measured correlation? Why?

(a.)

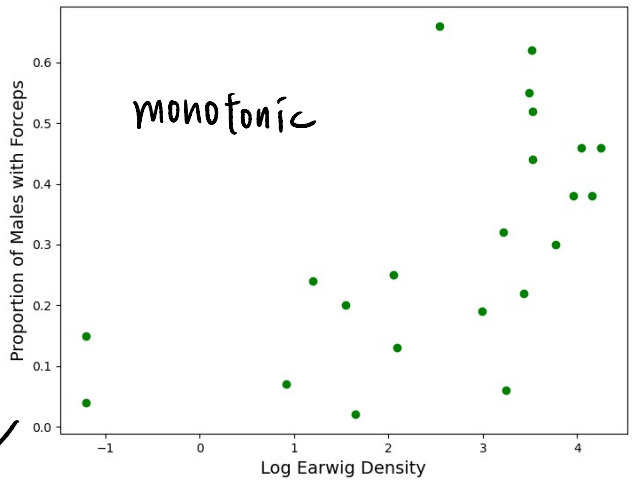




⇒ Not bivariate distribution



Log Transform →



(b). (c.)

Spearman's correlation: 0.661
p-Value: 0.001
} scipy.stats.spearmanr()

After transformation, we can see that there is a positive relationship between log-transformed earwig density and the proportion of males with forceps.
(p-Value < 0.05 ⇒ reject Ho that there is no relationship between two variables)

Ho: There is ($\rho_s = 0$) no relationship between earwig density and the proportion of males with forceps.

Ha: There is ($\rho_s \neq 0$) a relationship between earwig density and the proportion of males with forceps.

t-statistic: $\frac{\rho_s\sqrt{n-2}}{\sqrt{1-\rho_s^2}}$ (ρ_s: Spearman's rank correlation, n: sample size)

$= 3.944 > 0.425 \Rightarrow$ Reject Ho

Conclusion: Based on the analysis, we can reject Ho and conclude that there is a correlation between earwig density and proportion of males with forceps.

(d). If sampling errors are large, it may introduce additional noise into the measurements. ⇒ The true correlation would be stronger.

Weakens the observed relationship

Code for problem 2-4: https://github.com/Annie2305/HW2-4.git