Data Science Hw2

學號：b08502141　　　　姓名：石旻翰　　　　系級：電機四

1 Problem1

1.1 What is the unbiased estimator with the lowest variance that you can construct from a linear combination of θ_1 and θ_2, and what's its variance?

Suppose that

$$\theta' = a\theta_1 + (1-a)\theta_2 \tag{1.1}$$

Next,

$$\text{Var}(\theta') = a^2 \cdot 1 + (1-a)^2 \cdot 2 + 2a(1-a) \cdot \frac{1}{4} \tag{1.2}$$

Last, find the minimum of $\text{Var}(\theta')$, thus

$$\text{Var}(\theta') = 0.775, \text{when a } = 0.7 \tag{1.3}$$

1.2 What is the unbiased estimator with the lowest variance that you can construct from a linear combination of θ_3 and θ_4, and what's its variance?

Suppose that

$$\theta' = b\theta_3 + (1-b)\theta_4 \tag{1.4}$$

Next,

$$\text{Var}(\theta') = b^2 \cdot 1 + (1-b)^2 \cdot 2 + 2b(1-b) \cdot \frac{3}{4} \tag{1.5}$$

Last, find the minimum of $\text{Var}(\theta')$, thus

$$\text{Var}(\theta') = 0.958, \text{when b } = \frac{5}{6} \tag{1.6}$$

2 Problem2

2.1 Show that Q = X(1) - θ is a pivotal quantity

$$\begin{aligned}
F_Q(x) &= P(X(1) - \theta \leq x) = 1 - P(X(1) - \theta > x) \\
&= 1 - P(\bigcap_{i=1}^{n} \{X_i > \theta + x\}) \\
&= 1 - \prod_{i=1}^{n} P(X_i > \theta + x) \\
&= 1 - \prod_{i=1}^{n} \int_{\theta+x}^{\infty} e^{-(t-\theta)} dt \\
&= 1 - \prod_{i=1}^{n} e^{-x} \\
&= 1 - e^{-nx}, \text{is not a function of } \theta
\end{aligned}$$

2.2 Use this pivotal quantity find a 100(1-α)% confidence interval for θ

$$\begin{aligned} 1-\alpha &= P(a \le Q \le b) \\ &= P(a \le X(1)-\theta \le b) \\ &= P(X(1)-b \le \theta \le X(1)-a) \end{aligned}$$

From 2.1, we know that $F_Q(x) = 1 - e^{-nx}$, so

$$1 - e^{-na} = \frac{\alpha}{2} \Rightarrow a = -\frac{1}{n}\ln\left(1-\frac{\alpha}{2}\right)$$
$$1 - e^{-nb} = 1-\frac{\alpha}{2} \Rightarrow b = -\frac{1}{n}\ln\left(\frac{\alpha}{2}\right)$$
$$\Rightarrow P(X(1)-b \le \theta \le X(1)-a) = P\left(X(1)+\frac{1}{n}\ln\left(\frac{\alpha}{2}\right) \le \theta \le X(1)+\frac{1}{n}\ln\left(1-\frac{\alpha}{2}\right)\right)$$

3 Problem3

3.1 What are the probabilities of Type I error?

$$\text{Type I error} = P(\text{reject}|H_0) = P(X > 0.92|H_0 : \theta = 1)$$
$$= \text{P(X > 0.92) = 1 - P(X} \le 0.92) = 0.08$$

3.2 What are the probabilities of Type II error?

$$\text{Type II error} = P(\text{not reject}|H_a) = P(X \le 0.92|H_0 : \theta = 2)$$
$$= \text{P(X} \le 0.92) = 0.46$$

4 Problem4

4.1 Compute the significance level α of the test

To find the value of α, we first calculate λ, which is 8 * 0.5 = 4. And find the column of $\lambda = 4$, and we know that $\alpha = 0.0511$.

4.2 Write out the formula of the power function $\beta(\lambda)$ of this test

First,

$$\text{If} X_i \sim \text{Poisson}(\lambda), \text{then} \sum_{i=1}^{n} X_i \sim \text{Poisson}(n\lambda) \qquad [4.1]$$

Thus, the power function is,

$$\beta(\lambda) = 1 - P_{H_1}[\sum_{i=1}^{8} X_i \ge 8] = 1 - \sum_{y=0}^{7} \frac{e^{-8\lambda}(8\lambda)^y}{y!} \qquad [4.2]$$

And we can verify that

$$\text{When } \lambda = 0.5, \beta(\lambda) = \beta(0.5) = 1 - \sum_{y=0}^{7} \frac{e^{-4}4^y}{y!} = 0.0511 \qquad [4.3]$$

5 Problem5

5.1 Show your calculation steps: Code Implementation

You can run the code called lda.py to verify the output.

```
lda.py > ...
import numpy as np
import sys
# {(5, 3), (3, 5), (3, 4), (4, 5), (4, 7), (5, 6)}
# {(9, 10), (7, 7), (8, 5), (8, 8), (7, 2), (10, 8)}
data = [
    (5, 3, 0), (3, 5, 0), (3, 4, 0), (4, 5, 0), (4, 7, 0), (5, 6, 0),
    (9, 10, 1), (7, 7, 1), (8, 5, 1), (8, 8, 1), (7, 2, 1), (10, 8, 1)
    ]
sample, class1, class2 = [], [], []
for d in data:
    point = [d[0], d[1]]
    sample.append(point)
    if d[2] == 0:
        class1.append(point)
    else:
        class2.append(point)
print("Sample variance is ", np.var(sample, ddof=1))
class1, class2 = np.array(class1), np.array(class2)
u1 = np.mean(class1, axis = 0).reshape(2, 1)
u2 = np.mean(class2, axis = 0).reshape(2, 1)
s = np.matmul((u1 - u2), (u1 - u2).T)
s1 = np.zeros((2,2))
s2 = np.zeros((2,2))
for point in class1:
    s1 += np.matmul((point.reshape(2, 1)-u1),
    (point.reshape(2, 1)-u1).T)
for point in class2:
    s2 += np.matmul((point.reshape(2, 1)-u2),
    (point.reshape(2, 1)-u2).T)

sw = s1 + s2
a = np.matmul(np.linalg.inv(sw), s)
eigen_values, eigen_vectors = np.linalg.eig(a)
print("Max Eigen valud ", eigen_values[0] )
print("Normalize Eigen vector ", eigen_vectors[:, 0])
```

Figure 1: Code for LDA

5.2 Perform LDA and find out the optimal projection vector (normalized to unit length), and its corresponding eigenvalue

Run the code in Sec 5.1 and the output is shown as following:
Eigenvalue 5.212392829306312
Normalize Eigenvector [0.99130435 -0.13158907]

References

[1] https://www.math.arizona.edu/~jwatkins/N_unbiased.pdf

[2] https://zhuanlan.zhihu.com/p/372269493