e-Post Graduate Diploma in Advanced Business Analytics

Categorical Data Analysis: Assignment 1

Full marks: 110

Due date : EOD, Sunday, August 13, 2023

1. **[10 marks]** Identify whether each of the following variables are categorical (if so, binary, nominal or ordinal) or quantitative (if so, discrete or continuous) :

   (a) Number of publications of an IIMA faculty this year:

   (b) Whether work-from-home should be encouraged (yes, definitely; yes, probably; no, probably not; no, definitely not) :

   (c) Commando units of the Indian defence force (Garuda, Para, Ghatak, NSG, Marcos, SFF):

   (d) Number of hours per week an e-PGD participant studies outside of class:

   (e) Genre of movies in Netflix (Thriller, Romance, Horror, SciFi, Drama):

   (f) Anemic status of an individual (non-anemic, mildly anemic, moderately anemic, severely anemic):

   (g) Average amount you spend per month in online purchases:

   (h) Whether examples related to Covid-19 should be included in course materials:

   (i) Number of online courses you completed during lockdown:

   (j) Average number of hours per day you spend working from home nowadays:

   For the following problems, please attach the relevant `R` codes and outputs.

2. **[20 marks]** Suppose the probability of a head in a flip of a coin is π. Suppose the coin is flipped twice and Y heads are obtained.

   (a) **[3 marks]** Assuming $\pi = .50$, specify the probabilities for the possible values of Y and evaluate its mean and standard deviation.

   (b) **[2 marks]** Sketch the likelihood function of π given that you observe one head in the two flips.

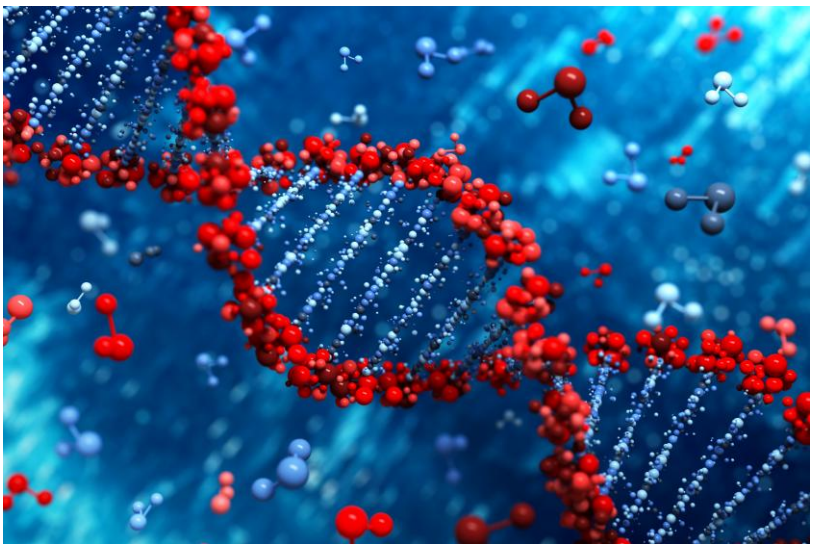
   (c) **[2 marks]** Using the plotted likelihood function above, determine the maximum likelihood estimate of π.

(d) **[2+1=3 marks]** Suppose that both tosses turn up tails. Find the corresponding maximum likelihood estimate of π. Does this estimate seem reasonable ? Justify.

(e) **[2+2=4 marks]** Find the probabilities of the possible values of Y when π equals i) 0.60 and ii) 0.40.

(f) **[2+2+2=6 marks]** Suppose the coin is tossed 100 times. Assuming $\pi = 0.40$, find the probability that the number of heads will be i) at least 40; ii) at most 65 and iii) between 45 and 65.

3. **[10 marks]** Genotypes AA, Aa and aa occur with probabilities (π_1, π_2, π_3). For $n = 3$ independent observations, the observed frequencies are (y_1, y_2, y_3).



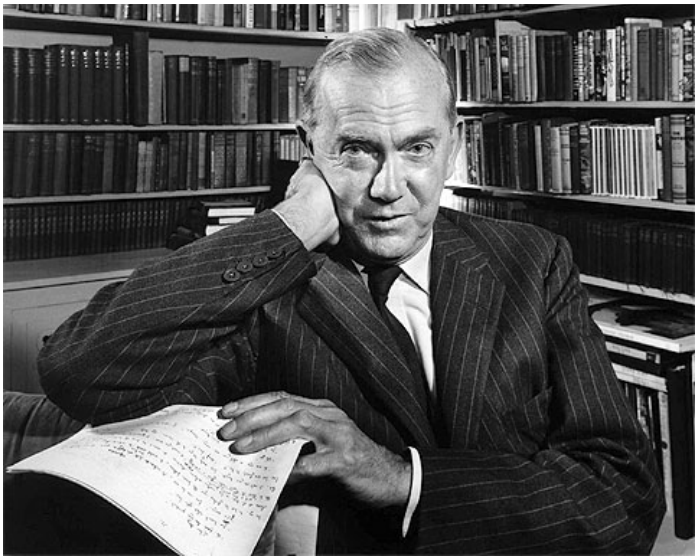
(a) **[2+1=3 marks]** Explain how you can determine y_3 from data on y_1 and y_2. Based on this what can you conclude about the dimensionality of the multinomial distribution ?

(b) **[3 marks]** Show the set of all possible observations (y_1, y_2, y_3) with $n = 3$.

(c) **[2 marks]** Suppose $(\pi_1, \pi_2, \pi_3) = (.25, .50, .25)$. Find the probability that $(y_1, y_2, y_3) = (1, 2, 0)$.

(d) **[2 marks]** For the above setup, find the probability distribution of y_1.

4. **[10 marks]** In his autobiography *A Sort of Life*, British author Graham Greene described a period of severe mental depression during which he played Russian Roulette.



This game consists of putting a bullet in one of the six chambers of a pistol, spinning the chambers to select one at random, and then firing the pistol once at one's head.

(a) **[2 marks]** Greene played this game six times, and was lucky that none of them resulted in a bullet firing. Find the probability of this outcome.

(b) **[2+2+2=6 marks]** Suppose Greene played the above 30 times in a row. Find the probability that i) at least 5 of the games will result in bullet fires ii) at most 10 of the games will result in bullet and iii) between 5 to 10 of the games will result in bullet fires.

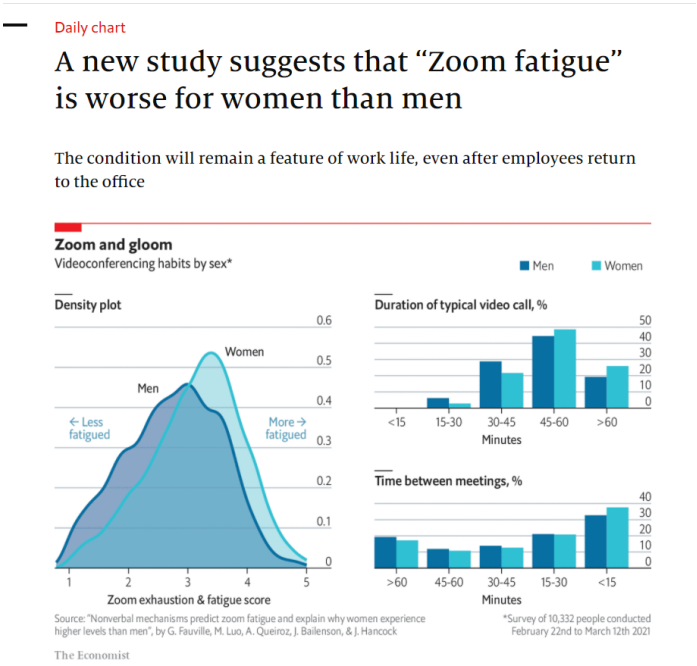
(c) **[2 marks]** Suppose he had kept playing this game until the bullet fires. Let Y denote the number of the game on which the bullet fires. Explain why the probability of the outcome y equals

$$f(y) = (5/6)^{(y-1)}(1/6), \quad \text{for} \quad y = 1, 2, 3, ...$$

5. **[60 marks]** **[Zoom fatigue]** Thanks to programs like ePGD-ABA, the lives of working professionals who would like to upskill themselves mainly revolve around Zoom nowadays. However, continued exposure to online mode of communication often leads to elevated levels of stress or *Zoom fatigue* as the following news article, (ref: *The Economist*, April 17th, 2021) depicts.

"Having endured endless virtual meetings over the past year, many workers are unsurprisingly complaining about "Zoom fatigue". Video-conferencing can be exhausting. Having to stay within the camera's gaze leaves limbs stiff and bottoms sore. Looking at your own face on

screen can be bad for self-esteem. And trying to communicate without all the usual visual cues (not least because of time-lags) adds to the "cognitive load" for already stressed-out employees."



This article is based on a study carried out by researchers at the University of Gothenburg in Sweden. A major takeaway of this study was that Zoom fatigue affects women more than males. Having said that, since the study participants were predominantly natives of the European countries, the findings may or may not apply to Indian working professionals.

[**Part A**]: Suppose you would like to figure out whether there is a similar gender-differential in Zoom fatigue between Indian working professionals as well i.e whether a higher percentage of female working professionals in India experience Zoom fatigue compared to their male counterparts. Towards that end, you mimic the aforementioned study on a random and representative sample of 2850 male and 2765 female working professionals from various corporations in India. Of them, 1815 males and 1895 females revealed symptoms of Zoom fatigue.

(a) [**1 mark**] State the research question for the above study.

(b) [**1+1=2 marks**] Identify the possible explanatory and response variables.

(c) [**5 marks**] The aforementioned study reported that about 63% of working professionals in Europe, across both genders reportedly suffered from Zoom fatigue. Carry out a likelihood ratio test at $\alpha = .05$ to verify whether the corresponding proportion for the Indian workforce is larger than the European estimate.

(d) [**1 mark**] Create a 2×2 contingency table cross classifying the above data across the categories of the response and explanatory variables you have identified in (2).

(e) [**1×5=5 marks**] Calculate the following:

i. The estimated probability that a randomly chosen working professional suffers from Zoom fatigue.

ii. The estimated probability that a randomly chosen female working professional suffers from Zoom fatigue.

iii. The estimated probability that a randomly chosen male working professional does not suffer from Zoom fatigue.

iv. The estimated probability that given a randomly chosen working professional is female, she will not suffer from Zoom fatigue.

v. The estimated probability that given a randomly chosen working professional is male, he will suffer from Zoom fatigue.

(f) **[4 marks]** Calculate the expected cell counts of the contingency table you created in (3) under the assumption of independence between the response and predictor variables.

(g) **[1+1+1=3 marks]** Estimate the conditional probabilities that a male and female working professional would suffer from Zoom fatigue and show that both of these are equal to the marginal probability of having zoom fatigue under the assumption of independence.

(h) **[1+2+2+1=6 marks]** Carry out a Pearsonian Chi-square test at $\alpha = .05$ to test whether gender has a significant association with susceptibility to Zoom fatigue. In doing so, specify the relevant hypotheses, calculate the test statistic, p-value and state your conclusion in the context of the problem.

(i) **[2+1=3 marks]** Calculate the difference in the estimated probabilities of suffering from Zoom fatigue across the genders. Interpret the same.

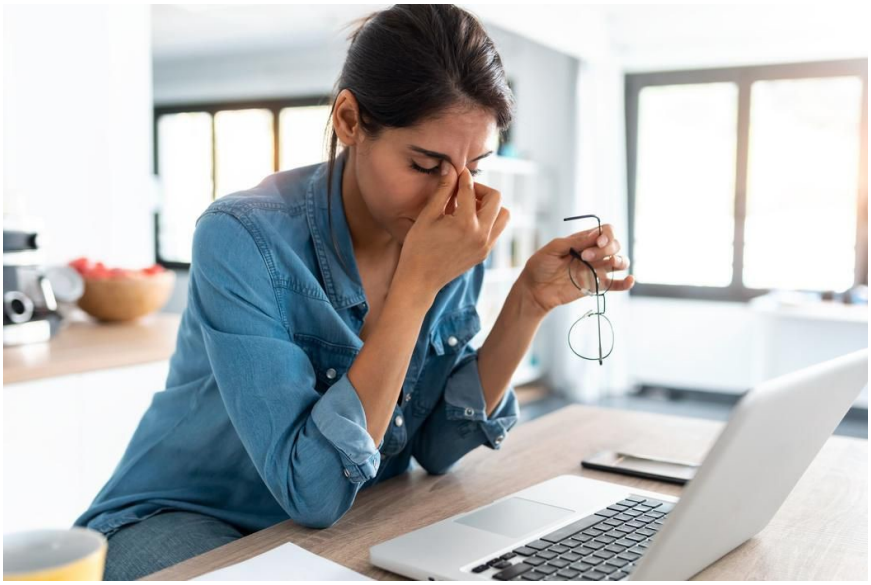
(j) **[2+1=3 marks]** Calculate the relative risk of being afflicted by Zoom fatigue between males and females. Interpret the same.

(k) **[3+1=4 marks]** Calculate the odds of suffering from Zoom fatigue for males and females and accordingly the odds ratio. Which gender seems to be more susceptible to Zoom fatigue ?

(l) **[2+1+1=4 marks]** Based on the above odds ratio value, calculate the change in odds of suffering from Zoom fatigue between males and females. Does this value corroborate the conclusion you had drawn in (10) above ? Justify. (*you are <u>not</u> supposed to use the conditional proportion values in the table).*

(m) **[3+1+1=5 marks]** Calculate a 95% confidence interval of the true population odds ratio of suffering from Zoom fatigue between males and females. Show all the relevant steps and interpret the interval in the context of the problem. Does this interval indicate any significant association between gender and proneness to Zoom fatigue ? Justify.

[Part B]: Suppose the true proportion of male and female Indian working professionals who are affected by Zoom fatigue in the population are p_1 and p_2 respectively. Accordingly answer the following questions.



(a) **[2 marks]** What would be a reasonable set of hypotheses to test whether susceptibility to Zoom fatigue depends on gender ?

(b) **[2+1=3 marks]** Evaluate the test statistic value. What is its null distribution ?

(c) **[2 marks]** Determine the p-value corresponding to the above test statistic.

(d) **[2 marks]** Based on the above p-value, you will (select the correct option/s)

i. Reject H_0 at $\alpha = 0.01$ but not at $\alpha = 0.05$.

ii. Reject H_0 at $\alpha = 0.01$ and 0.05 but not at $\alpha = 0.1$.

iii. Reject H_0 at $\alpha = 0.05$ and 0.1 but not at $\alpha = 0.01$.

iv. Reject H_0 at $\alpha = 0.1$ but not at $\alpha = 0.01, .05$.

v. Fail to reject H_0 at all the above α values.

vi. Reject H_0 at all the above α values.

(e) **[2 marks]** Based on your answer above, what would you conclude about the association between gender and proneness to Zoom fatigue ? Justify.

(f) **[3 marks]** Construct a 95% confidence interval of the difference in the proportions of male and female working professionals who are affected by Zoom fatigue ?

(g) **[1 mark]** Based on the above interval, what can you conclude about the significance of association between gender and susceptibility to Zoom fatigue ? Justify.

(h) **[1 mark]** Suppose your friend performed a similar survey in her organization. However, she could only collect data from a random sample of 210 female employees and 285 male employees. Everything else remaining the same, the 95% confidence interval of $p_1 - p_2$ will be

i. Narrower than the interval obtained in (6) above.

ii. Wider than the interval obtained in (6) above.

iii. Same as the interval obtained in (6) above.

iv. Need more information.

(i) **[2+1=3 marks]** Suppose, out of the 210 females whom your friend surveyed, all but 5 revealed that they suffer from Zoom fatigue. Accordingly, calculate a 95% *score confidence interval* of the true proportion of female working professionals in her organisation who suffer from Zoom fatigue. Interpret the same.