CSE 544 (Spring 2023)
Probability and Statistics for Data Science

Practice Mid-term 2

(6 questions, 33 points total)

I agree that engaging in dishonest behavior during the exam will result in a score of 0.
Dishonest behavior includes copying from other students, referring to any form of notes, conversing with other students without the permission of the instructor, etc.
By taking this exam, I acknowledge and agree to the above terms.

Please write your name here → ____

For instructor's use only.

Q1) 6 points:

Q2) 6 points:

Q3) 6 points:

Q4) 5 points:

Q5) 4 points:

Q6) 6 points:

Total (out of 33):

Q1) **(Total 6 points)**

Consider a distribution that takes value 3 with probability x and 0 with probability (1-x). You are given i.i.d. sample data D = {0, 3, 0, 0}.

(a) Find $\hat{x}_{MME}$. Show all your steps for a generic i.i.d. dataset D = {X_1, X_2, …, X_n}, and only at the end substitute for values from D and report your final answer as a number. (2 points)

(b) Find $\widehat{se}(\hat{x}_{MME})$. Show all your steps for a generic i.i.d. dataset D = {X_1, X_2, …, X_n}, and only at the end substitute for values from D and report your final answer as a number. (4 points)

(a) $\hat{\alpha}_1 = \alpha_1(x) \Rightarrow \bar{X} = 3.x + 0(1-x) = 3x$

$\Rightarrow \hat{x} = \frac{1}{3}\bar{X} = \frac{1}{3}\left(\frac{3}{4}\right) = \underline{\underline{0.25}}$

(b) $se = \sqrt{Var(\hat{x})} = \sqrt{Var\left(\frac{\bar{X}}{3}\right)} = \sqrt{\frac{1}{9} Var\left(\frac{\Sigma X_i}{n}\right)}$

$\overset{Lov}{\underset{iid}{=}} \sqrt{\frac{1}{9} \cdot \frac{Var(X)}{n}}$

$X \begin{cases} \cancel{x}\,3 & x \\ 0 & (1-x) \end{cases}$

$= \sqrt{\frac{1}{\cancel{9}} \cdot \frac{\cancel{9}x(1-x)}{n}} = \sqrt{\frac{x(1-x)}{n}}$

$E[X^2] = 9x \Rightarrow Var = 9x - (3x)^2$
$= 9x(1-x)$

$\hat{se} = \sqrt{\frac{\hat{x}(1-\hat{x})}{n}} = \sqrt{\frac{0.25 \times 0.75}{4}} = \sqrt{\frac{1}{4} \times \frac{3}{4} \times \frac{1}{4}} = \frac{\sqrt{3}}{8}$

Q2) **(Total 6 points)**

Consider a new test that is developed to detect the flu virus. In a population of 100 patients, 60 of them did not have flu and 40 of them had flu. The test was able to correctly predict the presence of flu in 30 patients but falsely predicted the presence of flu in 20 patients.

(a) Draw the 2 X 2 truth table for the above data using the same format, axes, and matrix entries as in class. Use the null as in class for medical testing. Hint: entries in each cell should be some positive integers. (2 points)

(b) What is the probability of Type-I error for this test and data? (2 points)

(c) What is the probability of False Negative for this test and data? (2 points)

ground	test says not sick Accept H_0	test says sick Reject H_0	
H_0 true	40 TN	FP 20 (Type I)	60
H_0 false	10 FN (Type II)	TP 30	40

(b) $Pr(\text{Type I error}) = Pr(\text{Reject } H_0 \mid H_0 \text{ true}) = \frac{20}{60} = 1/3$

(c) $Pr(FN) = Pr(\text{Accept } H_0 \mid H_0 \text{ false}) = \frac{10}{40} = 1/4$

Q3) **(Total 6 points)**

Consider an election with three parties: A, B, and C. Assume that out of 1000 sampled people, 600 were from urban communities and 400 were from rural communities. Of the 600 urban sampled people, 450 voted for party A, 50 for party B, and the remaining for party C. Of the 400 rural sampled people, 50 voted for party A, 0 voted for party B, and the remaining voted for party C.

(a) Draw the 2 X 3 table for this data with urban/rural as the rows and A, B, C as the columns. (2 points)

(b) Compute, numerically, the Q_{obs} metric for this example. Show all expected values for each of the 6 cells clearly. (4 points)

(a)

	A	B	C	
urban	300 / 450	30 / 50	270 / 100	600
rural	200 / 50	20 / 0	180 / 350	400
	500	50	450	1000
	1/2	1/20	9/20	

(b) $$Q_{obs} = \sum_{r,c} \frac{(E_{r,c} - O_{r,c})^2}{E_{r,c}} = \frac{(450-300)^2}{300} + \frac{(50-30)^2}{30} + \frac{(270-100)^2}{270} + \frac{(200-50)^2}{200} + \frac{(20)^2}{20} + \frac{(350-180)^2}{180}$$

Q4) **(Total 5 points)**

Let D={X_1, X_2, …, X_n} be a set of i.i.d. samples from a Uniform(0, θ) distribution, where θ is an unknown value. Let the prior for θ be some distribution W with pdf proportional to 1/θ. Find a posterior (1-α) interval for θ with all constants derived. Show all your steps clearly.

$$0 \le D \le \theta \Rightarrow \theta \ge \max\{X_1, \dots, X_n\} = m$$

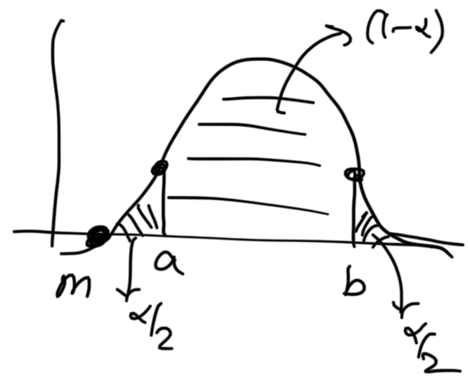
$$\Rightarrow \theta \in [m, \infty)$$

posterior $\propto \frac{1}{\theta^n}\cdot\frac{1}{\theta} = \theta^{-(n+1)} \Rightarrow$ posterior $= C\cdot\theta^{-(n+1)}$

$$\int_m^\infty C\cdot\theta^{-(n+1)}\cdot d\theta = 1 \Rightarrow C\cdot\frac{\theta^{-n}}{-n}\Big|_m^\infty = \frac{C}{n}\,\theta^{-n}\Big|_\infty^m = \frac{C}{n\cdot m^n}$$

$$\Rightarrow C = n\cdot m^n$$

$$\Rightarrow f(\theta|D) = n\cdot m^n\cdot\theta^{-(n+1)}$$



$$\boxed{\int_m^a f(\theta|D)\cdot d\theta = \alpha/2}$$

$$\Rightarrow \cancel{n}\cdot m^n\,\frac{\theta^{-n}}{\cancel{n}}\Big|_a^m = \alpha/2 \Rightarrow m^n\left(m^{-n} - a^{-n}\right) = \alpha/2$$

$$\Rightarrow 1-\left(\frac{m}{a}\right)^n = \alpha/2 \Rightarrow \frac{m}{a} = \sqrt[n]{1-\alpha/2} \Rightarrow a = \frac{m}{\sqrt[n]{1-\alpha/2}}$$

$$\int_b^\infty f(\theta|D)\cdot d\theta = \alpha/2 \qquad [a, b]$$

Q5) **(Total 4 points)**

Consider a simple linear regression estimation problem with no intercept term, that is, $\hat{Y}_i = \hat{\beta} X_i$. Let the objective to be minimized be S3 = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^3$. Let the dataset be {(1, 1); (2, 0); (0, 0)}. Recall that we order the data as (Y_i, X_i). Assume that the regression is applicable, and all assumptions are met. Use the OLS technique to determine the value of $\hat{\beta}$ for the given data and compute the MAPE error over the given dataset. You do not have to check for the 2[nd] derivative condition.

$$S3 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^3 = \sum_{i=1}^{n}(Y_i - \hat{\beta}X_i)^3$$

$$\frac{dS3}{d\beta} = 0 = \sum_{i=1}^{n} 3(Y_i - \hat{\beta}X_i)^2 \cdot (-X_i)$$

$$\Rightarrow \sum_{i=1}^{n} X_i(Y_i - \hat{\beta}X_i)^2 = 0$$

$$\Rightarrow \sum X_i(Y_i^2 - 2\hat{\beta}X_iY_i + \hat{\beta}^2X_i^2) = 0$$

$$\Rightarrow \sum \overset{1}{(X_iY_i^2)} - 2\hat{\beta}\overset{1}{\sum(X_i^2Y_i)} + \hat{\beta}^2 \cdot \overset{1}{\sum(X_i^3)} = 0$$

$$\Rightarrow 1 - 2\hat{\beta} + \hat{\beta}^2 = 0 \Rightarrow (\hat{\beta} - 1)^2 = 0 \Rightarrow \hat{\beta} = 1$$

$$\hat{Y}_i = X_i$$

$\hat{Y}_i$	Y_i X_i
1	1, 1
0	2, 0
0	0, 0

$$MAPE = \frac{\sum \left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100}{n} = \frac{1}{3}\left(\frac{2}{2} \times 100\right) = 33.33\%$$

Q6) **(Total 6 points)**

Let $\{X_1, X_2, \ldots, X_n\}$ be i.i.d. from Normal(μ_1, σ_1^2) and $\{Y_1, Y_2, \ldots, Y_m\}$ be i.i.d. from Normal(μ_2, σ_2^2). Also suppose *X's* and *Y's* are independent, and μ_1, σ_1^2, μ_2, σ_2^2 are unknown. Let S_x and S_Y be the sample standard deviations of the two populations. Assume that *n* and *m* are large. Let H_0: $\mu_1 \geq \mu_2$ be the null hypothesis and H_1: $\mu_1 < \mu_2$ be the alternate hypothesis. Consider the T statistic for the unpaired T test, as in class, with $\delta > 0$ being the critical value ($t_{n-1,\alpha}$).

(a) Show that the probability of Type-2 error is given by $1 - \Phi\left(-\delta - \frac{\mu_1 - \mu_2}{\sqrt{\frac{Sx^2}{n} + \frac{Sy^2}{m}}}\right)$. (4 points)

(b) Derive the p-value for the test. (2 points)

(a) Type-2 error = $Pr(\text{accept } H_0 \mid H_0 \text{ false})$

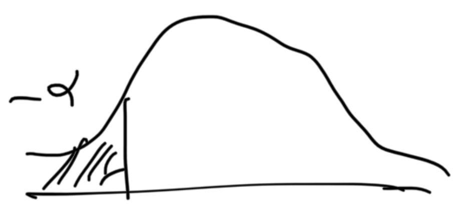
$\mu_1 < \mu_2$

$$\boxed{T = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}}$$

Reject if $T < -\delta$

$\Rightarrow$ Accept if $T \geq -\delta$

$Pr(T \geq -\delta \mid \mu_1 < \mu_2)$



$\bar{X} \sim Normal\left(\mu_1, \frac{\sigma_1^2}{n}\right)$ $\quad \bar{Y} \sim Normal\left(\mu_2, \frac{\sigma_2^2}{m}\right)$

$\bar{X} - \bar{Y} \sim Normal\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$

C

$$T \sim Normal\left(\frac{\mu_1 - \mu_2}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}, \frac{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}{\frac{S_x^2}{n} + \frac{S_y^2}{m}}\right) = Normal\left(\frac{\mu_1 - \mu_2}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}, 1\right)$$

$$Pr(T \geq -\delta) = Pr(\underbrace{T - C}_{Z} \geq -\delta - C) = 1 - Pr(Z < -\delta - c)$$

$$= 1 - \Phi(-\delta - c)$$

(extra page; intentionally left blank; use as needed)

$$\frac{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

Sample var is consistent

When n is large, sample var. $= \sigma_1^2$

$$\longrightarrow \frac{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} = 1$$

(b) p-value : $\Pr\Big(T < \underbrace{\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}}_{T_{obs}}\Big)$

$\Pr(T < T_{obs})$

$$c = \frac{\mu_1 - \mu_2}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

$T - c \approx Z$

$$\Pr(\underbrace{T - c}_{Z} < T_{obs} - c) = \Phi(T_{obs} - c)$$

$$= \Phi\left(\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}\right)$$

$H_0: \mu_1 \geq \mu_2$

H_0 true $=$ $\mu_1 = \mu_2$ $\quad \Phi\left(\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right) = \Phi(T_{obs})$

(extra page; intentionally left blank; use as needed)