408 Midterm 1 Solutions[1]

1. Question 1

TRUE/FALSE

(a)The negation of the statement
"There exists an integer j such that $j^3 - j < 7$" is:
"For every integer j, we have $j^3 - j \geq 7$."

TRUE, by the rules of negation, "There exists" is negated to"For every," and the inequality $<$ is negated to $\geq$.

(b) Let $\mathbf{P}$ be the uniform probability law on $[0,1]$. Let $x_1, x_2, \ldots \in [0,1]$ be a countable set of distinct points. Then

$$\mathbf{P}\left(\cup_{n=1}^{\infty}\{x_n\}\right) = 0.$$

TRUE. By the definition of $\mathbf{P}$, $\mathbf{P}(\{x_n\}) = 0$ for all $n \geq 1$. So, from Axiom (ii) for probability laws, $\mathbf{P}\left(\cup_{n=1}^{\infty}\{x_n\}\right) = \sum_{n=1}^{\infty} \mathbf{P}(\{x_n\}) = 0$.

(c) Let $X_1, \ldots, X_n$ be i.i.d random variables drawn from a family of probability density functions $\{f_\theta \colon \theta \in \mathbf{R}\}$ where $f_\theta \colon \mathbf{R} \to [0, \infty)$ for all $\theta \in \mathbf{R}$. Then there must exist some integer $k \geq 1$, $\exists$ some function $t \colon \mathbf{R}^n \to \mathbf{R}^k$ and there exists some statistic $Y = t(X_1, \ldots, X_n)$ such that Y is a sufficient statistic for θ.

TRUE. the statistic $(X_1, \ldots, X_n)$ is always sufficient for θ.

(d) Let $X_1, \ldots, X_8$ be i.i.d Gaussian distributed random variables. Define $W := \sum_{i=1}^{8} X_i$ and let $Z := \sqrt{\sum_{i=1}^{8}(X_i - W/8)^2}$. Then W and Z are independent.

TRUE. This follows from Proposition 3.7 in the notes.

(e) Let $Y_1, Y_2, \ldots$ be a sequence of estimators that are consistent for θ, where θ lies in a parameter space Θ. That is, $Y_1, Y_2, \ldots$ converges in probability to the constant random variable θ. Then $\mathbf{E}_\theta Y_n = \theta$ for all $n \geq 1$ and for all $\theta \in \Theta$.

FALSE. Consistency is just a statement about the limit as $n \to \infty$ of the sequence of estimators. For example, if could occur that $Y_1 = \infty$ or $Y_1 = 0$, while the sequence of estimators could still be consistent.

2. Question 2

Let $Y_1, Y_2, \ldots$ be random variables such that $\sqrt{n}Y_n$ converges in distribution to a mean zero Gaussian random variable with variance 5 as $n \to \infty$. Let

$$f(t) := (t+1)^3, \qquad \forall\, t \in \mathbf{R}$$

Show that, as $n \to \infty$, the random variables

$$\sqrt{n}(f(Y_n) - f(0))$$

converge in distribution to a random variable Z, and then compute $\mathbf{E}Z^2$.

Solution. From the first order delta method, $\sqrt{n}(f(Y_n) - f(0))$ converges in distribution to a mean zero Gaussian random variable with variance $5 \cdot (f'(0))^2 = 5 \cdot 3^2 = 45$. So, $\mathbf{E}Z^2 = 45$.

[1]September 27, 2021, © 2021 Steven Heilman, All Rights Reserved.

3. Question 3

Let θ be a an unknown real parameter, and suppose a random variable X has PDF

$$f(x) := \begin{cases} 1 & \text{, if } \theta \leq x \leq \theta + 1 \\ 0 & \text{, otherwise.} \end{cases}$$

- Find a method of moments estimator for θ.
- Find a method of moments estimator for θ^2.

Solution. We have

$$\mathbf{E}X = \int_{\mathbf{R}} xf(x)dx = \int_{\theta}^{\theta+1} xdx = (1/2)[(\theta+1)^2 - \theta^2] = \theta + 1/2.$$

That is, $\theta = \mathbf{E}X - 1/2$. So, a method of moments estimator for θ is

$$-1/2 + \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We have

$$\mathbf{E}X^2 = \int_{\mathbf{R}} x^2 f(x)dx = \int_{\theta}^{\theta+1} x^2 dx = (1/3)((\theta+1)^3 - \theta^3) = \theta^2 + \theta + 1/3.$$

That is,

$$\theta^2 = \mathbf{E}X^2 - \theta - 1/3 = \mathbf{E}X^2 - (\theta + 1/2) + 1/6 = \mathbf{E}X^2 - \mathbf{E}X + 1/6$$

So, a method of moments estimator for θ^2 is

$$1/6 + \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Alternatively, since $\theta^2 = (\mathbf{E}X - 1/2)^2$, another method of moments estimator for θ^2 is

$$\Big(-1/2 + \frac{1}{n}\sum_{i=1}^{n} X_i\Big)^2.$$

4. Question 4

Let $X_1, \ldots, X_n$ be a random sample of size n from a Bernoulli distribution with parameter $0 < \theta < 1$. (That is, $\mathbf{P}(X_1 = 1) = \theta$ and $\mathbf{P}(X_1 = 0) = 1 - \theta$, with θ unknown.)

Show that $Y := X_1 + \cdots + X_n$ is a sufficient statistic for θ.

Solution 1. Let $x_1, \ldots, x_n \in \{0, 1\}$ and let $0 \leq y \leq n$ be an integer. Then Y has a binomial distribution with parameters n and θ. We may assume that $y = x_1 + \cdots + x_n$, otherwise there is nothing to show. Then,

$$\mathbf{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \,|\, Y = y) = \frac{\mathbf{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n), Y = y)}{\mathbf{P}(Y = y)}$$

$$= \frac{\mathbf{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n))}{\mathbf{P}(Y = y)} = \frac{\prod_{i=1}^{n} \mathbf{P}(X_i = x_i)}{\mathbf{P}(Y = y)} = \frac{\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{y}\theta^y(1-\theta)^{n-y}}$$

$$= \frac{\theta^y(1-\theta)^{n-y}}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} = \frac{1}{\binom{n}{y}} = \frac{1}{\binom{n}{x_1 + \cdots + x_n}}.$$

Since the last expression does not depend on θ, Y is sufficient for θ.

Solution 2. Let $x_1, \ldots, x_n \in \{0, 1\}$. Then

$$\mathbf{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n)) = \mathbf{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n), Y = y)$$

$$= \mathbf{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n)) = \prod_{i=1}^{n} \mathbf{P}(X_i = x_i), \qquad \text{by independence}$$

$$= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = (1-\theta)^n \cdot [\theta/(1-\theta)]^{\sum_{i=1}^{n} x_i}$$

That is, we can write the joint distribution of $X_1, \ldots, X_n$ in the form $g_\theta(t(x)) \cdot h(x)$ for all $x = (x_1, \ldots, x_n) \in \{0, 1\}^n$, where

$$g_\theta(w) := (1-\theta)^n \cdot [\theta/(1-\theta)]^w, \qquad h(x) := 1, \qquad t(x) := \sum_{i=1}^{n} x_i.$$

So, from the factorization theorem, $t(X) = X_1 + \cdots + X_n$ is sufficient for θ.

5. Question 5

Consider a population of 30,000 people, where half of them are given a vaccine for a disease. Suppose all 30,000 people are exposed to a virus causing the disease. We observe that 90 of the unvaccinated people catch the disease, while 5 of the vaccinated people catch the disease.

Consider the following statement:

"If we have a population of 30,000 people exposed to the virus, with half of them vaccinated, then the number of infections of vaccinated people, divided by the number of infections of unvaccinated people, is less than 15/100."

Is the statement true with greater than 90% certainty? Justify your answer.

(Assume that each person's ability to catch the disease is independent of each other person's ability to catch the disease.)

(Hint: the estimated probability of a vaccinated person getting the disease is $5/15,000$, and the estimated probability of an unvaccinated person getting the disease is $90/15,000$.)

(Hint: use the Central Limit Theorem. If Z is a standard Gaussian, then $\mathbf{P}(|Z| \leq 2) \approx .9545$. Also, $\sqrt{5} \approx 2.23$, $\sqrt{90} \approx 9.5$.)

Solution. Let X_i be the indicator random variable which is 1 if the i^{th} vaccinated person catches the disease and 0 if not, for all $i \in \{1, 2, \ldots, 15,000\}$. Let Y_i be the indicator random variable which is 1 if the i^{th} unvaccinated person catches the disease and 0 if not, for all $i \in \{1, 2, \ldots, 15,000\}$. Then we are assuming the $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ are i.i.d. with $\mathbf{P}(X_1 = 1) = p = 5/15,000$ and thus $\mathbf{E}[X_1] = p$ $\text{var}(X_1) = p(1-p)$. Also, $\mathbf{P}(Y_1 = 1) = q = 90/15,000$ and thus $\mathbf{E}[Y_1] = q$ $\text{var}(Y_1) = q(1-q)$. The statement can be written as

$$\frac{X_1 + \cdots + X_{15000}}{Y_1 + \cdots + Y_{15000}} < .15. \qquad (*)$$

Then by the central limit theorem, we have

$$\mathbf{P}\left(-2 \leq \frac{X_1 + \cdots + X_{15,000} - 15,000p}{\sqrt{15,000p(1-p)}} \leq 2\right) \approx .9545$$

Since $15000p = 5$ and $150000p(1-p) = 5(1-p) \approx 5$, we have

$$\mathbf{P}\left(5 - 2\sqrt{5} \leq X_1 + \cdots + X_{15,000} \leq 5 + 2\sqrt{5}\right) \approx .9545$$

That is,

$$\mathbf{P}\left(0 \leq X_1 + \cdots + X_{15,000} \leq 10\right) \approx .9545$$

Meanwhile,

$$\mathbf{P}\left(-2 \leq \frac{Y_1 + \cdots + Y_{15,000} - 15,000q}{\sqrt{15,000q(1-q)}} \leq 2\right) \approx .9545$$

Since $15000q = 90$ and $150000q(1-q) = 90(1-q) \approx 90$, we have

$$\mathbf{P}\left(90 - 2\sqrt{90} \leq Y_1 + \cdots + Y_{15,000} \leq 90 + 2\sqrt{90}\right) \approx .9545$$

That is,

$$\mathbf{P}\left(71 \leq Y_1 + \cdots + Y_{15,000} \leq 109\right) \approx .9545$$

So, with at least 95.45% certainty, the number of vaccinated people with the disease is at most 10. Also, with at least 95.45% certainty, the number of unvaccinated people with the disease is at least 71. So, with at least 90% certainty, we have

$$\frac{X_1 + \cdots + X_{15000}}{Y_1 + \cdots + Y_{15000}} \leq \frac{10}{71} < .15.$$

So, the statement is true with at least 90% certainty.