Name: HUAN Q. BUI

Statistics 482 Spring 2020
Exam 1

Wednesday, 1 April 2020

- This is an open-note and open-book exam.
- All work must be your own. You may not give or receive any kind of aid, either verbally, visually, or otherwise, during this exam. No other sources may be consulted, except as specified above.
- The exam has 90 possible points. There are 6 questions and 10 pages, including this cover page.
- A standard normal table and information concerning common distributions' densities, expectations, variances, and moment-generating functions is available on Moodle.
- You have 4 hours to complete the exam so plan your time accordingly. I have included the possible points next to each problem.
- Some questions are more difficult than others, and the questions may not be in order of difficulty.
- Don't spend too much time on any one question; if you get stuck, go on and try another part.
- Whenever possible, show your work and explain your reasoning. In case you make a mistake, I can more easily give you partial credit if you explain your steps.
- Please type answers directly into questions, or provide a document with answers neatly written. Use the file submission option in Moodle to upload your document.

Question 1 (12 points total)

A. (3 points) Let Y be a random variable that has an F-distribution with v_1 numerator and v_2 denominator degrees of freedom. Find the distribution of $U = 1/Y$.

B. (9 points) Suppose that Y_1, Y_2, Y_3, Y_4, and Y_5, are a random sample from a normal population with mean 0 and standard deviation 1. Also let $\bar{Y} = \frac{1}{5}\sum_{i=1}^{5} Y_i$, and let Y_6 be another independent observation from the same distribution. What is the distribution of each of the following? Explain your rationale briefly.

a. $\boldsymbol{\sum_{i=1}^{5} Y_i^2}$

b. $\boldsymbol{\sum_{i=1}^{5} (Y_i - \bar{Y})^2}$

c. $\boldsymbol{\sum_{i=1}^{5} (Y_i - \bar{Y})^2 + Y_6^2}$

Question 2 (18 points total)

A random sample of 37 second graders who participated in sports had manual dexterity scores with mean 32.19 and standard deviation 4.34. An independent sample of 37 second graders who did not participate in sports had manual dexterity scores with mean 31.68 and standard deviation 4.56.

A. **(6 points)** Perform an appropriate hypothesis test to see if there is sufficient evidence to indicate that second graders who participated in sports have a higher mean dexterity score. Use $\alpha = 0.05$. You may assume that the standard deviations given above are *population* values and that the distribution of dexterity scores is approximately normal. Make sure to report all necessary components: hypotheses, test statistic, p-value, and conclusion.

Question 2 continued…

B. (6 points) For the test in part A, calculate the power when $\mu_{sports} - \mu_{no\ sports} = 3$.

C. (6 points) Write an algorithm (not formal R code, but just the process) to perform a bootstrap-based test for the scenario described in part A.

Question 3 (20 points total)

Let Y be a single randomly drawn observation from a distribution with density given by:

$$f_Y(y) = \frac{2(\theta - y)}{\theta^2}; \qquad 0 < y < \theta$$

A. (8 points) Show that $U = Y/\theta$ is a pivotal quantity.

B. (4 points) Use the pivotal quantity in part A to find a 90% lower confidence limit for θ. Note that you should start with the expression: $P(U < a) = 0.90$, where a is the lower confidence limit. (Hint: you may need the quadratic formula: $y = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

Question 3 continued…

C. **(8 points)** If you wanted to create a random sample of observations having the distribution of *Y* above (assume a known, but fixed, θ) from *n* random uniform variables on the interval (0,1). Describe the algorithm you would use if you wanted to utilize inverse transform sampling. Be as specific as possible.

Question 4 (20 points)

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample drawn from a distribution with density given by,

$$\boldsymbol{f_Y(y)} = \begin{cases} \left(\dfrac{\mathbf{1}}{\boldsymbol{\theta^2}}\right)\boldsymbol{ye^{-y/\theta}}, & \boldsymbol{y > 0} \\ \mathbf{0}, & \textbf{elsewhere} \end{cases}$$

Note that *Y* has a gamma distribution with parameters $\alpha = 2$ and $\beta = \theta$.

A. **(6 points)** Find the maximum likelihood estimator, $\hat{\theta}$, of θ.

B. **(2 points)** What is the maximum likelihood estimator of $V(Y_i)$?

C. **(6 points)** Find expected value, $E(\hat{\theta})$, and variance, $V(\hat{\theta})$, of the estimator you found in part A.

Question 4 continued…

D. (6 points) Is $\hat{\theta}$ from part (A) a consistent estimator of θ? Support your statement.

Question 5

Consider a sequence of random variables $\{X_n\}$, where each has density given by:

$$f_{X_n}(x) = \begin{cases} \left(1 - \frac{x}{n}\right)^{n-1}; & 0 \le x \le n \\ 0; & \text{elsewhere} \end{cases}$$

A. **(4 points)** Find the cumulative distribution function of X_n.

B. **(6 points)** What is the limiting distribution of this sequence of random variables (i.e., what does this sequence converge to in distribution)?

Question 6 (10 points total)

For each of the statements below, indicate whether it is true or false.

1. We are unable to calculate the power of a hypothesis test unless we have both a simple null hypothesis and a simple alternative hypothesis.

True **False**

2. The sum of *n* i.i.d random exponential(2) random variables has a exponential (2n) distribution.

True **False**

3. We can compare the p-value to alpha because they are both conditional probabilities that are both conditional on the null being true.

True **False**

4. A bootstrap sample is taken with replacement because we are simulating the population by pretending that we have an infinite number of copies of our original sample.

True **False**

5. One of the main advantages of the accept-reject algorithm (acceptance sampling), as compared with inverse transform sampling, is that you do not need to know the density function of the random sample you are trying to generate.

True **False**

Question 1.

(a) $Y \sim F(v_1, v_2)$. So Y has the form

$$Y = \frac{A/v_1}{B/v_2}$$

where $A \sim \chi^2(v_1)$ and $B \sim \chi^2(v_2)$. So,

$$U = \frac{1}{Y} = \frac{B/v_2}{A/v_1} \sim F(v_2, v_1).$$

(b)

- Because for $Y_i \sim \mathcal{N}(0, 1)$, $Y_i^2 \sim \chi^2(1)$ and $\sum^n Y_i^2 \sim \chi^2(n)$, we have that

$$\sum_{i=1}^{5} Y_i^2 \sim \chi^2(5).$$

- We know that

$$\frac{n-1}{\sigma^2} S^2 = \frac{n-1}{\sigma^2} \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sim \chi^2(n-1),$$

so

$$\sum_{i=1}^{5} (Y_i - \bar{Y})^2 = \frac{5-1}{1^2} \frac{1}{5-1} \sum_{i=1}^{5} (Y_i - \bar{Y})^2 \sim \chi^2(5-1) = \chi^2(4).$$

- From (a) and (b) we know that

$$\sum_{i=1}^{5} (Y_i - \bar{Y})^2 \sim \chi^2(4), \quad Y_6^2 \sim \chi^2(1).$$

Since these are independent, we have a theorem that says

$$\sum_{i=1}^{5} (Y_i - \bar{Y})^2 + Y_6^2 \sim \chi^2(4+1) = \chi^2(5).$$

Question 2.

(a) We have $H_0 : \mu_{\text{sports}} = \mu_{\text{no sports}}$, $H_a : \mu_{\text{sports}} > \mu_{\text{no sports}}$, and $\alpha = 0.05$. The test statistic is

$$z = \frac{(\bar{x}_{\text{sports}} - \bar{x}_{\text{no sports}}) - 0}{\sqrt{\frac{\sigma_s^2}{n_s} + \frac{\sigma_{ns}^2}{n_{ns}}}} = \frac{32.19 - 31.68}{\sqrt{\frac{4.34^2}{37} + \frac{4.56^2}{37}}} \approx 0.4927.$$

All conditions are met/assumed. The p-value is $1 - 0.6879 \approx 0.312 > 0.05 = \alpha$. So, there is not enough evidence to reject H_0, i.e., there is not enough evidence to indicate that second graders who participated in sports have a higher dexterity score.

(b) Power is the probability of rejecting H_0 provided that H_a is true. To find power we want to find the critical value for the test statistic. At $\alpha = 0.05$, single-tail, $z_c = 1.645$. So,

$$\begin{aligned} \text{Power} &= P(z \geq z_c | \Delta\mu = 3) \\ &= P(z' = z - z_c \geq -1.355 | \Delta\mu = 0), \quad \text{(shifting)} \\ &= 0.912. \end{aligned}$$

(c) Bootstrap algorithm to test the scenario in part (a):

$$\begin{aligned} H_0 &: \mu_{ns} = \mu_s \\ H_a &: \mu_s > \mu_{ns}. \end{aligned}$$

- Calculate the observed $\bar{X}_s - \bar{X}_{ns} = \bar{\Delta}$.
- Combine all observations into one sample, called Z.
- Take a bootstrap sample from Z of size $n_1 = 37.$, then calculate $\bar{X^*}_s$.
- Take a bootstrap sample from Z of size $n_2 = 37.$, then calculate $\bar{X^*}_{ns}$.
- Calculate $\bar{X^*}_s - \bar{X^*}_{ns} = \Delta^*$.
- Repeat steps (3)-(5) and get a distribution for Δ^*.
- The p-value is the number of $\Delta^* > 0$ divided by the number of bootstraps.
- Compare this p-value to $\alpha = 0.05$ and conclude.

Question 3. Y is a r.v. with

$$f_Y(y) = \frac{2(\theta - y)}{\theta}, \quad 0 < y < \theta.$$

(a) $U = Y/\theta$ is a function of the sample measurement, and of only one unknown parameter θ. Further, with $h^{-1}(u) = y = u\theta$

$$f_U(u) = f_Y(h^{-1}(u))\left|\partial_u h^{-1}(u)\right| = \frac{2(\theta - u\theta)}{\theta^2} \cdot \theta = 2(1-u), \quad u \in (0,1)$$

does not depend on θ or any unknown parameter. So $U = Y/\theta$ is a pivotal quantity.

(b) We want to find an a such that $P(U < a) = 0.90$.

$$P(U < a) = \int_0^a 2(1-u)\, du = 2a - a^2 = 0.90.$$

Using the quadratic formula, we have $a = 0.6837$ or $a = 1.316$. We reject the latter because of the condition $a \in (0,1)$. So, $a = 0.6837$. So,

$$P(U < 0.6837) = P(Y/\theta < 0.6837) = P(\theta > Y/0.6837) = 0.90.$$

(c) We want to use the inverse transform to generate r.v. distributed the same as Y. We know $f_Y(y)$. So, $F_Y(y)$ is

$$F_Y(y) = \int_0^y \frac{2(\theta - y')}{\theta^2}\, dy' = \frac{2y}{\theta} - \frac{y^2}{\theta^2}.$$

Let $u = T^{-1}(y) = T^{-1}(T(u)) = F_Y(y)$ with $u \sim U(0,1)$. Then solving for y using the quadratic formula gives

$$y = \theta - \theta\sqrt{1-u}, \quad \text{or} \quad \theta + \theta\sqrt{1-u}.$$

Since $y \in (0,\theta)$, we reject the second solution. With this, we generate a sample of uniform $u \sim U(0,1)$, then let $y = \theta - \theta\sqrt{1-u} \sim \text{pdf}(Y)$.

Question 4.

(a) To find the mle of θ, $\hat{\theta}$, we write down the log likelihood function:

$$l(\theta) = \ln(\mathcal{L}(\theta)) = \ln\left(\prod_{i=1}^{n} \frac{1}{\theta^2} y_i e^{-y_i/\theta}\right) \tag{1}$$

$$= \ln\left(\frac{1}{\theta^{2n}} e^{-\sum_{i=1}^{n} y_i/\theta} \prod_{i=1}^{n} y_i\right) \tag{2}$$

$$= -2n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} y_i + \ln\prod_{i=1}^{n} y_i. \tag{3}$$

Taking the derivative w.r.t. θ and setting it to zero gives

$$\frac{-2n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} y_1 = 0 \iff \hat{\theta} = \frac{1}{2n}\sum_{i=1}^{n} y_i = \frac{\bar{y}}{2}. \tag{4}$$

(b) The mle of $V(Y_i)$, with $Y_i \sim \Gamma(2, \theta)$, is

$$\text{mle}(V(Y_i)) = \text{mle}(2\theta^2) = 2\hat{\theta}^2 = 2\left(\frac{\bar{y}}{2}\right)^2 = \frac{\bar{y}^2}{2}.$$

where we have used the invariance property of mle in the second equality.

(c)

$$E(\hat{\theta}) = E\left[\frac{1}{2n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{2n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{2n}\sum_{i=1}^{n} 2\theta = \frac{n\theta}{n} = \theta$$

$$V(\hat{\theta}) = V\left[\frac{1}{2n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{(2n)^2}\sum_{i=1}^{n} V(Y_i) = \frac{n(2\theta^2)}{(2n)^2} = \frac{\theta^2}{2n}.$$

(d) Since $E(\hat{\theta}) = \theta$, $\hat{\theta}$ is an unbiased estimator for θ. Also,

$$\lim_{n\to\infty} V(\hat{\theta}) = \lim_{n\to\infty} \frac{\theta^2}{2n} = 0.$$

By the handy theorem, $\hat{\theta} \xrightarrow{P} \theta$, i.e., $\hat{\theta}$ is a consistent estimator for θ.

Question 5.

(a)

$$F_{X_n}(x) = \int_0^x \left(1 - \frac{x'}{n}\right)^{n-1} dx' = -\frac{n}{n}\left(1 - \frac{x'}{n}\right)^n \Bigg|_0^x = 1 - \left(1 - \frac{x}{n}\right)^n, \quad 0 \le x \le n.$$

(b)

$$\lim_{n\to\infty} F_{X_n}(x) = \lim_{n\to\infty}\left[1 - \left(1 - \frac{x}{n}\right)^n\right] = 1 - e^{-x}, \quad 0 \le x.$$

This is the cdf of the Exp(1). So, $X_n \xrightarrow{D} X \sim$ Exp(1).

Question 6.

(a) **True.** Power is the probability of rejecting H_0 when H_a is true. To actually calculate the power we need both H_0 and H_a to be in simple form (assigning a specific value to the parameter), or else the power is indeterminate.

(b) **False.** $\text{Exp}(\beta) = \Gamma(1, \beta)$. We also know that if iid $X_i \sim \Gamma(1, \beta)$ then $\sum^n X_i \sim \Gamma(n, \beta)$. But $\Gamma(n, \beta) \neq \text{Exp}(n\beta)$. So this statement is false.

(c) **True.** This is true by definition of the p-value and the significance level α. We can make such a comparison because these are conditional on the same thing. It is unfair to compare probabilities conditional on different things.

(d) **True.**

(e) **False.** We <u>do</u> need to know the density function to run the accept-reject algorithm.