Stat 88 Spring 2021 **FINAL EXAM Solutions**

1. The density of a random variable X is given by

$$f(x) = \begin{cases} x & 0 \leq x \leq 1 \\ 2-x & 1 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find $E[X]$. (2 points)

(b) Find the cdf of X. Write the following probability in terms of the cdf of X and then compute it: $P(X \leq \frac{3}{2} \mid X > \frac{1}{2})$. (2+1+2 points)

(c) Let $X_1, X_2, \ldots, X_{100}$ be i.i.d random variables with the same distribution as X. What is the distribution of the **number** of X_i's such that $\frac{1}{2} < X_i \leq \frac{3}{2}$? Write an expression for the chance that exactly 75 of the X_i's are below 0.5. (2+1 points)

Solution

Part a. $E[X] = \int_0^1 x{\cdot}x\,dx + \int_1^2 x{\cdot}(2-x)\,dx = 1$ (2 points)

Or you could draw the distribution and see it form a triangle symmetric around 1

Part b. CDF $F(x)$ will be piecewise.

$$F(x) = \begin{cases} 0 & x \leq 0 \\ \int_0^x s\,ds = \frac{x^2}{2} & 0 \leq x \leq 1 \\ \int_0^1 s\,ds + \int_1^x 2 - s\,ds = 2x - \frac{x^2}{2} - 1 & 1 \leq x \leq 2 \\ 1 & x \geq 1 \end{cases}$$

(2 points)

$$P(X \leq \frac{3}{2} \mid X > \frac{1}{2}) = \frac{P(\frac{1}{2} < X_i \leq \frac{3}{2})}{P(X > \frac{1}{2})} \qquad \text{(1 point)}$$

$$= \frac{F(\frac{3}{2}) - F(\frac{1}{2})}{1 - F(\frac{1}{2})} \qquad \text{(1 point)}$$

$$= \frac{7/8 - 1/8}{7/8} = 6/7 \qquad \text{(1 point)}$$

Part c.

If Y = the number of X_i's that lie between 1/2 and 3/2, then since $P(1/2 < X < 3/2) = 3/4$, $Y \sim \text{Bin}(100, 3/4)$.
Note that I have a typo: I meant to use the probability I have written here, and that they would have had to compute in Part(b), but in the question I asked "Write an expression for the chance that exactly 75 of the X_i's are below 0.5.". In this case $Y \sim \text{Bin}(100, p)$, where $p = P(X < 1/2) = 1/8$. In either case, please give them credit. We want:
$P(Y = 75) = \binom{100}{75} p^{75}(1-p)^{25}$. (2 points)

2. Office hours in Data 8 tend to be crowded before project deadlines. Nancy, who needs help on her project, makes the following observation. If there are **no more than 5 students** AND **no less than 5 staff members** on the queue, she can receive help within a minute.

Let X and Y be the number of students and staff members on the queue, respectively.
Nancy estimates the following quantities from her experiences: $E[X] = 9$, $SD(X) = 2$, $E[Y] = 2$, $SD(Y) = 2$.

Use the inequalities you have learned in Stat 88 to provide the best bounds for the probability that Nancy receives help within a minute.

You will do this by:

(a) Finding the best bounds for the chance that there are no more than 5 students. (3 points)

(b) Finding the best bounds for the chance that there are no fewer than 5 staff. (3 points)

(c) Now finding the best bounds for Nancy getting help within a minute. (4 points)

Solution

Part a)

- By Markov's inequality: $0 \le \mathbb{P}(X > 5) \le \mathbb{P}(X \ge 5) \le \frac{9}{5} > 1 \implies 0 \le (X \le 5) \le 1$ (1 point)
- By Chebyshev's inequality: $0 \le \mathbb{P}(X \le 5) \le \mathbb{P}(|X - E[X]| \ge 2 \cdot SD(X)) \le \frac{1}{4}$ (1 point)
- The best bounds for X are: $0 \le \mathbb{P}(X \le 5) \le \frac{1}{4}$ (1 point)

Part b)

- By Markov's inequality: $0 \le \mathbb{P}(Y \ge 5) \le \frac{2}{5}$ (1 point)
- By Chebyshev's inequality: $0 \le \mathbb{P}(Y \ge 5) \le \mathbb{P}(|Y - E[Y]| \ge \frac{3}{2} \cdot SD(Y)) \le \frac{4}{9}$ (1 point)
- The best bounds for Y are: $0 \le \mathbb{P}(Y \ge 5) \le min\{\frac{4}{9}, \frac{2}{5}\} = \frac{2}{5}$ (1 point)

Part c) The question does not make independence assumption about the number of students and the number of staff members. However, you can still get half the credit if you reach the correct answer under the independence assumption. **No double penalty if their answers to part a and part b are incorrect.**

- $\mathbb{P}$(Nancy receives help within a minute w/o indep assumption) $= \mathbb{P}(X \le 5, Y \ge 5)$
  - Upper bound: $\mathbb{P}(X \le 5, Y \ge 5) \le min\{\mathbb{P}(X \le 5), \mathbb{P}(Y \ge 5)\} \le min\{\frac{1}{4}, \frac{2}{5}\} = \frac{1}{4}$ (3 points)
  - Lower bound: $\mathbb{P}(X \le 5, Y \ge 5) \ge max\{\mathbb{P}(X \le 5) + \mathbb{P}(Y \ge 5) - 1, 0\} \ge 0$ (1 point)
- $\mathbb{P}$(Nancy receives help within a minute w/ indep assumption) $= \mathbb{P}(X \le 5, Y \ge 5) = \mathbb{P}(X \le 5)\mathbb{P}(Y \ge 5) \implies 0 \le \mathbb{P}(X \le 5)\mathbb{P}(Y \ge 5) \le \frac{1}{4} \cdot \frac{2}{5} = \frac{1}{10}$ (2 points)

3. In a population of 100 adults who are terrible at cooking, 40 take a comprehensive cooking course with a renowned chef. Among those who take the class, 30 adults experience an improvement in their cooking skills. Among those who do not take the class, 30 adults also experience an improvement in their cooking skills for other reasons. Test, at the 5% level, the hypothesis that this cooking class improved the skills of those who took it.

   Make sure to clearly state:
   - the null and alternative hypotheses, (1 point each)
   - the test statistic (1 point)
   - its distribution under the null hypothesis, (2 points)
   - write down the observed value of the test statistic (1 point)
   - compute the P-value (you will need to use python for this - or a calculator), (3 points)
   and make your conclusion. (1 point)

Solution

Let X be the number of improved adults among those who took the class. Then the hypothesis test is as follows:

(a) Null hypothesis: the class had no effect on cooking skills. (1 point)

(b) Alternative hypothesis: the class had a positive effect on (i.e., improved) cooking skills. (1 point)

(c) Let $T = X - \mathbb{E}_{H_0}[X]$. Then larger values of T provide stronger support for the alternative hypothesis and smaller values support the null hypothesis, as desired. Note that, under the null hypothesis, there is no difference between those who took the class and those who didn't. Hence the number of "improved" adults can be considered constant across the population, and we can assume the adults who improved in the class would have also improved without it. Full credit is also given if X is stated as the test statistic instead of T. (1 point)

Thus under the null hypothesis $X \sim Hg(N = 100, G = 60, n = 40)$. (2 points)
Hence $\mathbb{E}_{H_0}[X] = 40(\frac{60}{100}) = 24$, meaning our test statistic is $T = X - 24$. We observe $t^{obs} = 30 - 24 = 6$. If X is used as test statistic, the observed value is 30. (1 point)

(d) Calculate P-value.

$$\begin{aligned} P &= \mathbb{P}_{H_0}(T \geq t^{obs}) \\ &= \mathbb{P}_{H_0}(X - 24 \geq 6) \\ &= \mathbb{P}_{H_0}(X \geq 30) \\ &= \sum_{k=30}^{40} \frac{\binom{60}{k}\binom{40}{40-k}}{\binom{100}{40}} \\ &= 0.0103 \end{aligned}$$

(Where the numerical value is calculated using Python.) (3 points)

(e) Since $0.0103 < 0.05$, we reject the null hypothesis at the 5% level. **Full credit is given as long as the conclusion is consistent with your p-value.** (1 point)

4. You are interested in investigating when the Cal Falcons leave their nest on top of the Campanile. You know that in California, the **length** of a falcon's first flight (call this Y) follows an **exponential** distribution with an expectation equal to the **age** of the falcon when they have their first flight. The age that a falcon has their first flight is a continuous random variable (call this X), and whose pdf is given by:

$$f(x) = \begin{cases} 0 & x < 2, \\ \frac{32}{3x^3} & 2 \leq x \leq 4, \\ 0 & x > 4 \end{cases}$$

   (a) What is the expected length of a falcon's first flight? (4 points)

   (b) You are interested in the median age of a falcon when they have their first flight (that is, the median of the distribution of X). Since the length of a falcon's first flight (Y) depends on the age at which it makes this flight, you decide to use this length to estimate the median. Is this estimator unbiased? What does the bias of the median tell you about the shape of the distribution of X? (4+2 points)

Solution

For all three parts, let X = falcon's age at first flight and Y = length of falcon's first flight. Then, $Y \sim Exp(1/X)$

Part a) $E(Y) = E(E(Y|X)) = E(X)$ (2 points)

$E(Y) = \int_2^4 x \frac{32}{3x^3} dx = 8/3$ (2 points)

Part b) Let m = median of the distribution of X, then by the definition of the median, $F(m) = 1/2$. Note that

$F(m) = \int_2^m \frac{32}{3x^3} dx = 1/2$ (2 points)

$\frac{1}{m^2} = \frac{5}{32}$ so $m = \sqrt{\frac{32}{5}}$ (2 points)

$m < E(Y)$ so Y is **not** an unbiased estimator of m. (1 point)

Also, since $E(Y) = E(X) > m$, this indicates that the distribution of X is skewed-right. (1 point)

5. Nancy is interested in the average height of the undergraduate students at Berkeley. She guesses that the average height is **67 inches**. To test her guess, she would like to take a simple random sample of 100 students and use the sample mean ($\bar{X}_1$) to estimate the true average height of an undergraduate student at Berkeley.

   (a) Nancy observes that the average height of her sample is **66 inches** and the SD of the sample is **3 inches**. Construct a 95% confidence interval for the true average height of the an undergraduate student at Berkeley. Based on your confidence interval, do you think Nancy's guess of 67 inches is reasonable? (2+1 points)

   (b) Nancy then does some research and finds that the heights of undergraduate and graduate students at Berkeley both have mean **66 inches** and standard deviation **4 inches**. Suppose she wants to take another simple random sample of 100 graduate students, and denote the average height of this sample by $\bar{X}_2$. Assume that $\bar{X}_1$ and $\bar{X}_2$ are independent, and let $M = max\{\bar{X}_1, \bar{X}_2\}$. Find $P(M < 67)$. (4 points)

   (c) Suppose Nancy constructed two 95% confidence intervals for the true mean height of a student at Berkeley using two independent random samples of students. What is the probability that **neither** of these confidence intervals is successful at covering the true mean height? (3 points)

Solution

Part a.

95% CI: $66 \pm z \cdot 3/\sqrt{100} = 66 \pm 1.96 \times 3/10 = [65.412, 66.588]$. $z = 2$ is acceptable. (2 points)
Since 67 is not in the interval, Nancy's guess does not look reasonable. (1 point)

Part b. By CLT, both $\bar{X}_1$ and $\bar{X}_2$ are Normal$(66, \sigma = 4/\sqrt{100})$, or Normal$(66, \sigma = 0.4)$. Since M is the maximum of $\bar{X}_1$ and $\bar{X}_2$, $M < 67$ means both $\bar{X}_1$ and $\bar{X}_2$ are smaller than 67.

$$\begin{aligned}
P(M < 67) &= P(\bar{X}_1 < 67, \bar{X}_2 < 67) && \text{(1 point)}\\
&= P(\bar{X}_1 < 67) \cdot P(\bar{X}_2 < 67) && \text{(1 point)}\\
&= [P(\bar{X}_1 < 67)]^2 &&\\
&= [P(\frac{\bar{X}_1 - 66}{0.4} < \frac{67 - 66}{0.4})]^2 && \text{(1 point)}\\
&= [P(\bar{X}_1^* < 2.5)]^2 &&\\
&= [\Phi(2.5)]^2 && \text{(1 point)}
\end{aligned}$$

Part c. Each 95% CI is successful with a probability of 0.95. The chance that *neither* of them is successful is 0.05^2 since they are independently computed. (3 points)

6. Let Y_i be the first **year** that an individual i (selected randomly from the population) started working, and T_i be the "job tenure" of the random individual i, calculated in 2021, which is simply $2021 - Y_i$ or the **number of years** individual i has been working. Let W_i be the hourly **wage** of the random individual i. Let Z_{Y_i} be Y_i in standard units, Z_{T_i} be T_i in standard units, and Z_{W_i} be W_i in standard units. An economist estimates **two** linear regression models and the results are presented here:

$$\hat{W}_i = -0.88Y_i + 0.36$$
$$\hat{Z}_{W_i} = -0.44Z_{Y_i}$$

Another economist measures the job tenure data and tells you that $SD(T_i) = 2$.

(a) Find $SD(W_i)$. (2 points)

(b) Find $r(W_i, T_i)$. (2 points)

(c) Suppose another economist wants to predict Z_{W_i} using Z_{T_i}. They know that the scatter plot of the random pair (Z_{T_i}, Z_{W_i}) is football-shaped. For an individual with $Z_{T_i} = 2$, what is the estimated percentile rank for Z_{W_i}? You may write your answer in terms of $r(W_i, T_i)$. (2 points)

(d) Yet another economist comes along and tells them that these regressions are not worth the trouble, and they should just use the average value of W_i as a predictor, rather than regressing W_i on Y_i or T_i. Given that $\sqrt{1-r^2}$ is almost 90% what would you say to your colleague? (2 points)

(e) For those individuals whose standardized job tenure (Z_{T_i}) is at the 50th percentile rank, what percentile rank do you predict for their standardized wage (Z_{W_i})? Choose from one of the following and explain: (2 points)

(a) about 50% (b) less than 50% (c) more than 50%

Solution

Part a. By linear regression line formula, we have $r(W_i, Y_i) \cdot \frac{SD(W_i)}{SD(Y_i)} = -0.88$ and $r(W_i, Y_i) = r(Z_{W_i}, Z_{T_i}) = -0.44$. Thus, $SD(W_i) = 2SD(Y_i)$. (1 point) Note that $T_i = 2021 - Y_i$, which means $Y_i = 2021 - T_i$. Since $SD(T_i) = 2$, $SD(Y_i) = SD(2021 - T_i) = SD(T_i) = 2$. Hence, $SD(W_i) = 4$.(1 point)

Part b. Note that $T_i = 2021 - Y_i$. Thus, $r(W_i, T_i) = r(W_i, 2021 - Y_i) = -r(W_i, Y_i) = 0.44$. This uses the result proved in Chapter 11 Exercise 5. (2 points)

Part c. Using the linear regression results, we have $\hat{Z}_{W_i} = r(Z_{W_i}, Z_{T_i})Z_{T_i} = r(W_i, T_i)Z_{T_i} = 0.44 \times 2 = 0.88$. (1 point) Football-shaped diagram indicates we can assume these standard units are normally distributed. Thus, the estimated percentile rank is $\Phi(0.88) \approx 81\%$.(1 point)

Part d. If $\sqrt{1-r^2}$ is almost 90% there is some reduction in the size of the residuals, that is the the SD of residuals from regression is significantly less than the SD of the response, and therefore, the variation in the W_i wage is explained to some extent by its relationship with the year of beginning work. (2 points)

Part e. About 50% since the regression line goes through the point of averages. (2 points)

7. You and three colleagues from work have traveled to the city of Arrakeen for a conference on water conservation. All of you prefer to take individual taxis (there is no Lyft or Uber on Arrakeen) from the airport to your hotel, because of the pandemic, rather than travel together. And then you wonder how many taxis there are in this city. When the taxicabs arrive, each of you notices the serial number on your cab. Your taxi's serial number is **394**. Your colleagues have the serial numbers **31, 191**, and **278**, respectively. Suppose that N is the total number of taxis in the city fleet. Assuming that every taxi in the city fleet has a unique serial number (with the numbering beginning at 1), and that each taxi was picked uniformly at random from the fleet (of course, without replacement).

   You want to estimate N. How would you do it? One of your colleagues wants to use the sample mean. Another wants to use the sample median (for this sample, it is the average of 191 and 278) in this case. The third colleague wants to use the maximum (which is 394).

   Which one should you use? You will get points for:
   - computing the expected values of the sample mean, sample median, and sample maximum (5 points)
   - writing down the formulas for the unbiased estimators, (3 points)
   - the observed values of the estimators for this particular sample, and (1 points)
   - deciding which estimator you would use. (1 points)

   It might help to draw a stylized picture like in the text of dots, and mark these sampled values, and then proceed to compute the expected values that you will need.