**Stat 88 Spring 2021** **FINAL EXAM** **S. M. Stoyanov**

**NAME (FIRST LAST):** _____ **SID:** _____

**TIME AND CONDITIONS:**

- You have **150 minutes** to complete the exam and **30 minutes** to upload your submission to Gradescope. We consider this ample time and **no late submissions will be allowed**.
- If you have any technical issues, email your exam **immediately, within your allowed time** to stat88exams@gmail.com. We will **not** grade submissions that arrive more than 3 hours after you began your exam, nor anything that arrives after 10 pm on Tuesday, May 11 (even if you began your exam at 9:30 pm - it is your responsibility to begin your exam at an appropriate time).
- The exam is open-book, open-notes, open-Jupyter notebook, open-calculator, but **no other materials are allowed**. You are not allowed to look for answers on the internet, nor discuss any matter regarding the exam with anyone. You should write all your assumptions on your exam submission, and we will take them into consideration while grading.
- We **will not be available for answering questions on Piazza during the exam period**. This is because it is not possible for us to monitor Piazza overnight, and in order to ensure equitable treatment for all students, we will not monitor it at all. The Piazza site will be closed at 7 pm on Monday, when the exam period begins. Therefore, it is very important that you show all your work and state all your assumptions.
- You **must** begin **each QUESTION** of a question on a **separate** page. THIS IS IMPORTANT.
- You must select the correct page associated for each subpart of the question while submitting your exam on Gradescope. If we do not see your submitted answer to a problem, we are not going to search for it, but just give you a 0 for that problem.
- You may use a tablet to write your solutions.
- **At the top of the first page you are submitting, please write your full name and sign next to it.** This indicates that you have read and agreed to abide by the following Honor Code statement (you do not need to transcribe the statement):
  Data Science and the entire academic enterprise are based on one quality - integrity. We are all part of a community that doesnt fabricate evidence, doesnt fudge data, doesnt present other peoples work as our own, doesnt lie and cheat. You trust that we will treat you fairly and with respect. We trust that you will treat us and your fellow students fairly and with respect. Please read carefully the (slightly adapted) UC Berkeleys Honor Code below:
  **I certify that all solutions will be entirely my own and that I will not consult or share information with other people during the exam (including strangers on the internet). I promise I will act with honesty, integrity, and respect for others.**
- Please note that if we suspect you of cheating, we will report the incident to the Center of Student Conduct, and assign you an Incomplete grade for the course, pending case resolution. If we are *certain* that you have cheated, we will give you a 0 in the exam (and therefore an F in the course), *and* report you for academic misconduct.

**QUESTIONS AND ANSWERS**

- There are 7 questions for a total of 70 points. **Do not forget to put your name and signature at the top of your exam** - if this is missing we will take off 10% of your score.
- **Give brief explanations or show calculations in each question** unless the question says this is not required. You may use, without proof, any result proved or used in lecture, the textbook, and homework, unless the question asks for a proof.
- Please leave answers as **unsimplified arithmetic or algebraic expressions including finite sums** unless the question asks for a simplification.

**GRADING**

- The exam is worth **70** points. Each problem is worth 10 points.
- Please commit yourself to a **single, clearly marked** answer for each question. If you give multiple answers (such as both True and False) then please don't expect credit, even if the right answer is among those that you gave.
- **It is your responsibility to complete and submit the exam on time.** A late submission will not be accepted under any circumstances.

**FORMAT**

- **You must answer each QUESTION on a separate page for a (MINIMUM) total of 8 pages scanned.**
- **You must select the correct page** associated for each subpart of each question when submitting to Gradescope. If you do not, you will get a zero for that question on the exam.

---

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

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

(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: $P(X \le \frac{3}{2} \mid X > \frac{1}{2})$. (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 $\frac{1}{2} < X_i \le \frac{3}{2}$ ? Write an expression for the chance that exactly 75 of the $X_i$'s are below 0.5. (2+1 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)

---

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)

---

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)

---

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 Nancys 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)

---

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%

---

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.*