Take-Home Final Exam

DSE 210

Winter 2020

Guidelines for the Final Exam

1. Students must submit their solutions before 3/17/2020, 11:59 PM. Late submissions are not allowed.
2. The final has to be done individually. NO group work is allowed.
3. The solutions for the Final can be handwritten or typed. (Note: If the handwriting is illegible or if the pages are not scanned properly then the student shall receive 0 points for the respective question)
4. The solutions have to be uploaded to Gradescope in PDF format only. (Note: Do not forget to map question numbers and the pages containing the respective answers while uploading on Gradescope)
5. Some questions have the tag "***Programming Question***". Students are expected to implement the solution for the respective question in Python. PDF version of the source code has to be uploaded to Gradescope. (Note: Students should combine the theoretical solutions and Python source codes into one PDF file and then upload it to Gradescope)

DSE 210: Probability and statistics **Winter 2020**

Take-Home Final Part 1 — Sampling

1. (2 points) A box contains 9 red marbles and 1 blue marbles. Nine hundred random draws are made from this box, with replacement. What is distribution of the number of red marbles seen, roughly?

2. (5 points, 1 each) A dartboard is partitioned into 20 wedges of equal size, numbered 1 through 20. Half the wedges are painted red, and the other half are painted black. Suppose 100 darts are thrown at the board, and land at uniformly random locations on it.

   (a) Let X_i be the number of darts that fall in wedge i. What are $\mathbb{E}(X_i)$ and $\text{var}(X_i)$?

   (b) Using a normal approximation, give an upper bound on X_i that holds with 95% confidence.

   Let Z_r be the number of darts that fall on red wedges, let Z_b be the number of darts that fall on black wedges, and let $Z = |Z_r - Z_b|$ be the absolute value of their difference. We would like to get a 99% confidence interval for Z. To do this, define

   $$Y_i = \begin{cases} 1 & \text{if } i\text{th dart falls in red region} \\ -1 & \text{if } i\text{th dart falls in black region} \end{cases}$$

   and notice that $Z_r - Z_b$ can be written as $Y_1 + Y_2 + \cdots + Y_{100}$, the sum of independent random variables.

   (c) What are $\mathbb{E}(Y_i)$ and $\text{var}(Y_i)$?

   (d) Using the central limit theorem, we can assert that $Z_r - Z_b$ is approximately a normal distribution. What are the parameters of this distribution?

   (e) Give a 99% confidence interval for Z.

3. (2 points, 1 each) You have hired a polling agency to determine what fraction of San Diegans like sushi. Unknown to the agency, the actual fraction is exactly 0.5.

   The agency is going to poll a random subset of the population and return the observed fraction of sushi-lovers. How far off would you expect their estimate to be (i.e. what standard deviation) if:

   (a) they poll 100 people?

   (b) they poll 2500 people?

4. (3 points) In a certain city, there are 100,000 people age 18 to 24. A random sample of 500 of these people is drawn, of whom 194 turn out to be currently enrolled in college. Estimate the percentage of all persons age 18 to 24 in the city who are enrolled in college. Give a 95.5% confidence interval for your estimate.

5. (2 points) A survey research company uses random sampling to estimate the fraction of residents of Austin, Texas, who watch Spanish-language television. They are satisfied with the estimate they get using a sample size of 1,000 people.

   They then want to also estimate this fraction for Dallas, which has similar demographics to Austin, but twice the population. What sample size would be suitable for Dallas?

6. (2 points, 1 each) A box contains many pieces of papers with numbers on them. 100 random draws are made from the box, with replacement, and the sum of the draws is 297.
   (a) Can you estimate the average of the numbers in the box?
   (b) Can you give a confidence interval for your estimate, based on the information so far?

DSE 210: Probability and statistics **Winter 2020**

Take-Home Final Part 2 — Hypothesis testing

7. (3 points, 1 each) The Public Health Service studied the effects of smoking on health, in a large sample of representative households. For men and women in each age group, those who had never smoked were on average somewhat healthier than the current smokers, but the current smokers were on average much healthier than those who had recently stopped smoking.
   (a) Was this a controlled experiment or an observational study?
   (b) Why did they study men and women and the different age groups separately?
   (c) The conclusion seems to be that you shouldn't start smoking, but once you've started, you shouldn't stop. Comment.

8. (3 points, 1 each) In 10,000 tossings, a coin came up heads 5,400 times. Should we conclude that the coin is biased?
   (a) Formulate the null hypothesis and alternative hypothesis.
   (b) Compute the z statistic and the p-value.
   (c) What do you conclude?

9. (4 points) A die is rolled 100 times. The total number of spots is 368 instead of the expected 350. Can this be explained as chance variation, or is the die loaded?

10. (2 points) Other things being equal, which is better for the null hypothesis: a higher p-value or a lower p-value?

11. (4 points, 2 each) The National Household Survey on Drug Abuse was conducted in 1985 and 1992. In each year, a simple random sample of 700 people was used.
    (a) Among persons age 18 to 25, the percentage of marijuana users dropped from 21.9% to 11.0%. Is this difference real, or a chance variation?
    (b) Among persons age 18 to 25, the percentage of cigarette smokers dropped from 36.9% to 31.9%. Is this difference real, or a chance variation?

12. (4 points) A random sample of 1000 freshmen at public universities were asked how many hours they worked each week (for pay). The average number of hours turned out to be 12.2, with a standard deviation of 10.5. A similar survey at private universities had an average of 9.2 hours, with a standard deviation of 9.9. Is the difference between these two averages due to chance?

13. (4 points) A survey was conducted to determine the distribution of marital status by sex for persons age 25-29 in Wyoming. A random sample of 103 people was chosen, of whom 48 were men and 55 were women. The following results were obtained:

	Men	Women
Never married	43.8%	16.4%
Married	41.7%	70.9%
Widowed, divorced, separated	14.6%	12.7%

Are the distributions really different for men and women?

14. ***Programming Question:*** (10 points, 2.5 points each)
Consider the weather data that we used before. This time we will focus on Philadelphia and New York weather data. Similar to what we did before, we will again convert Kelvin to Fahrenheit. Assuming we are in March 2017, we will try to understand if the temperatures of this month are within known limits.

    (a) Filter March temperature records from all years, using all the data from all March records, calculate the mean and standard deviation. Do this separately for Philadelphia and New York. Later we will use these parameters as population parameters.

    (b) Filter the data for March 2017. Using this data calculate the sample means for Philadelphia and New York. We will use these sample means as our observations. Using these observations provide 99% confidence intervals for the true mean. Do this separately for Philadelphia and New York. We already know the true means from part a. Do the true means lie within the intervals?

    (c) Based on our prior knowledge we assert that March average temperatures for Philadelphia and New York are the same. Based on March 2017 temperature recordings we would like to test our prior knowledge. Conduct this test using hypothesis testing, please state your hypotheses as well. For population standard deviations use the standard deviations that you found in part a.

    (d) Repeat part c but this time assume that your sample sizes are 10. This means we have the same average March 2017 temperatures as part c. We will again use the population standard deviation from part a. However, we will use a different sample size this time.

DSE 210: Probability and statistics **Winter 2020**

Take-Home Final Part 3 — General Questions

15. (6 points, 3 each) A smart-phone consists of 10 major components, each independently faulty with probability $1 - p$. If any component is faulty, the phone is damaged. Five phones are manufactured independently in sequence. Find the probability that:

    (a) the last phone manufactured is the first damaged,

    (b) exactly two are damaged.

    Write your answers as powers of p and $1 - p$.

16. (6 points, 2 each) A fair coin with $P(heads) = 0.5$ and a biased coin with $P(heads) = 0.75$ are placed in an urn. One of the two coins is picked at random and tossed twice. Find the probability:

    (a) of observing two heads,

    (b) that the biased coin was picked if two heads are observed.

    (c) Qualitatively explain your answer to part (b) in 1-2 sentences.

17. (6 points, 2 each) Alice has 6 balls and Bob has 10. Each of them rolls an independent fair die and gives the other as many balls as their roll's outcome. For example, if Alice rolls 2 and Bob rolls 5, they will end up with 6 - 2 + 5 = 9 and 10 - 5 + 2 = 7 balls respectively. Find the probability that Alice ends up with:

    (a) strictly more balls than Bob,

    (b) the same number of balls as Bob,

    (c) strictly fewer balls than Bob.

18. (6 points, 2 each) n balls are tossed at random into n bins, so that each ball is equally likely to fall in any bin, and different balls are independent of each other. For example, for $n = 3$, balls 1, 2, and 3, may fall in bins 2, 2, and 1, respectively. Express each of the following in terms of n.

    (a) The probability that bin 1 is empty.

    (b) The expected value $E(X)$, where X is the number of empty bins, e.g., for the $n = 3$ illustration above, just bin 3 is empty, hence $X = 1$.

    (c) The probability that bins 1 and 2 are empty.

19. (5 points) Let X and Y be independent random variables with expectations 1 and 2, and variances 3 and 4, respectively. Find the variance of XY.

20. (3 points, 1 each) Which distribution would you prefer for the following tasks? and why?

    (a) The number of times you would win against Magnus Carlsen in chess when played n games, given that probability of you winning is p in each game.

(b) Number of shoots Kobe Bryant had to make until seeing his first goal. (Let p be the probability of securing a goal)

(c) To determine the probability of exactly 4 storms occurring in 2021 at Toronto, given that on an average 2 storms occur each year.

21. (5 points) Random variable X is distributed Poisson, and $P(X = 2) = P(X = 4)$. Find $P(X = 3)$.

22. (6 points, 2 each) A computer manufacturer produces 2000 chips, each with independent defect probability 0.001. Using the Poisson approximation for the number X of defective chip, find:

    (a) The Poisson parameter for X

    (b) $P(X > 1)$

    (c) $P(X \leq 3)$

23. (4 points) The quadratic function $f : \mathbb{R}^3 \longrightarrow \mathbb{R}$ given by

$$f(x) = 4x_1^2 + x_1x_2 + 9x_2^2 - x_2x_3 + 16x_3^2$$

    can be written in the form of x^TMx for some symmetric matrix M. What is M?

24. (3 points) If $M = \begin{bmatrix} 1 & 3 \\ -6 & -9 \end{bmatrix}$, then find M^{-1}

25. Bonus: (4 points, 2 each) An experienced fisherman says that salmon length has distribution $N(5, 1)$ and sea bass length has distribution $N(10, 4)$, where $N(\mu, \sigma^2)$ is univariate normal distribution. The fisherman also says that prior probabilities for catching a salmon and sea bass are $\mathrm{P}(salmon) = 2/3$, $P(bass) = 1/3$. Given this information:

    (a) What is the decision boundary for classifying a fish caught based on its length?

    (b) If a fish has length 7 units, what would your prediction be based on Bayes-optimal prediction?

    Probability density function of $N(\mu, \sigma^2)$:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$