Swarthmore Honors Examination 2021: Statistics
Kevin Ross (Cal Poly)

1. This is a closed-material exam. You may not refer to any books, notes, online sources, software, or any other resources, except for a calculator to do arithmetic.
2. Show your work and explain your reasoning in clearly identified steps, calculations, formulas, carefully chosen words, or well labeled graphs. Responses without sufficient or appropriate justification will not receive credit.
3. This is a three-hour exam. Budget your time wisely; some parts will be harder than others. It is better to give partial solutions to as many problems as possible rather than complete solutions to some problems but no solutions to others. In particular, don't spend time "perfecting" your solutions to the parts you're most confident in. There are 9 problems of varying length; you should make valiant attempts on at least 7 of them.
4. You may use common facts from statistics and probability without proof or citation, but be sure to provide a brief statement of the fact. ("The sample mean is the MLE of the parameter of a Poisson distribution." "The variance of a Uniform(a, b) distribution is $(b-a)^2/12$." "It follows from the Central Limit Theorem that..."etc.
5. If you cannot answer part of a problem, but need its answer for a later part, just make up a *reasonable* value and use it. For example "I can't do part (a), so I am going to assume that the answer is 0.2 so I can do part (b)." This applies to multiple steps within a single part and non-numerical answers as well. ("I'm not sure about the analysis but I'll assume we reject the null hypothesis to make my conclusion.") If you get stuck, do whatever you need to do to keep making progress (e.g., make stuff up).
6. There might be some parts that contain computations that you might not be able to do by hand or with a simple calculator. You can answer these parts by explaining how you would perform the computations. For example, you can provide commands you would use if you had access to software (e.g, "I would find the value of z using `qnorm(0.975, 0, 1)`) in R and then I would do $100 + 10z$"). You can also draw well labeled pictures representing how would you perform the computations. It's important that you demonstrate understanding of the principles behind the computations, even if you can't carry out them out. Even in situations that you can compute, numerical answers without sufficient or appropriate justification will not receive credit.
7. Carefully read the instructions for each problem and answer in the manner indicated.
8. A few questions ask "how would you use simulation?" For these questions
   - You need to describe the process in words, and not write actual code.
   - Your answer must include a hypothetical but detailed numerical example of what the results from one repetition of the simulation might look like.
   - You should not attempt to conduct the simulation or specify the results; you just need to describe the process you would follow in detail.
9. Write your solutions neatly on separate sheets of paper. Be sure to clearly label your responses (1a, 1b, 2, etc).

GOOD LUCK!!!

1. A statistician throws a dart at a dartboard. The dartboard can be represented as an xy-coordinate plane with the bullseye at the origin $(0, 0)$. Suppose that (X, Y) is the landing point of any particular dart, and X and Y are independent, each following a Normal distribution with mean 0 and *variance* θ. Then $R = \sqrt{X^2 + Y^2}$ is the distance between the landing point of the dart and the bullseye.

   The statistician throws n darts, independently, and measures the distances $R_1, \ldots, R_n$. We wish to use the R_i values to estimate θ.

   (a) Without doing calculus, explain why each R_i^2 has an Exponential distribution with rate parameter $\frac{1}{2\theta}$ (and mean 2θ).
   (b) Show that the probability density function (pdf) of each R_i is

$$f_R(r) = \frac{r}{\theta} \exp\left(-\frac{r^2}{2\theta}\right), \quad r > 0$$

   (c) Show that the maximum likelihood estimator of θ, based on $R_1, \ldots, R_n$, is

$$\hat{\theta} = \frac{1}{2n} \sum_{i=1}^{n} R_i^2$$

   (d) Show that the MLE $\hat{\theta}$ has variance

$$\mathrm{Var}(\hat{\theta}) = \frac{\theta^2}{n}$$

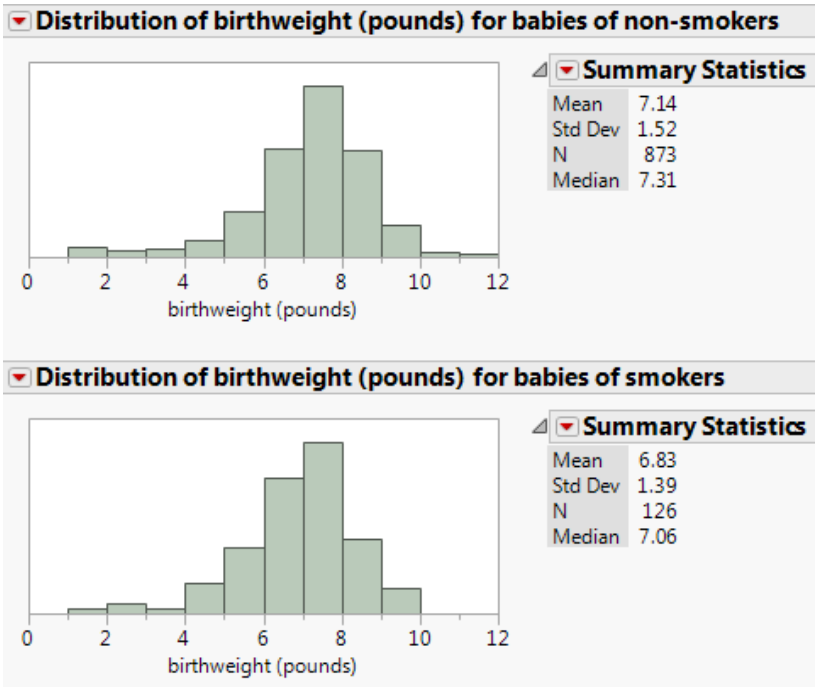
   (e) Show that $I_1(\theta)$, the Fisher information corresponding to a single observed R, is

$$I_1(\theta) = \frac{1}{\theta^2}$$

   (f) Is the MLE $\hat{\theta}$ the uniformly minimum unbiased estimator (UMVUE) of θ? Explain.
   (g) If $n = 4$ and the observed R values are 1.5, 0.3, 2.1, 0.7, find an *exact* 95% confidence interval for θ.

2. You have two friends, Fred and George, who are identical twins. They both like to play basketball, but Fred more so than George. Suppose that the probability that Fred successfully makes any particular free throw attempt is 0.9, and his attempts are independent. Similarly for George, but with a success probability of 0.7. One day while walking in the park, you see one of the twins — you can't tell which one — shooting free throws. Because you know that Fred likes basketball more than George, your prior probability that the shooter is Fred is 2/3. As you're walking you observe the shooter attempt 10 free throws, of which 8 are successful. You decide to watch one more attempt, and if the shooter makes the next attempt you'll shout "Fred!" and if he misses it you'll shout "George!". (Note: This is not necessarily a good strategy.) Find the probability that you'll call the shooter by the correct name.

3. The following summarizes data from a study involving birth weights (pounds) of babies and whether or not the baby's birth mother was a smoker.



An article written about the study features the headline "Babies born to smokers weigh significantly less than babies born to non-smokers." Is this an appropriate headline? Answer yes or no, and justify your answer with an appropriate analysis based on the information provided.

4. Let $X_1, X_2, \ldots$ be i.i.d. Poisson(10). Let Y_n be the product of the first n X_i's,

$$Y_n = \prod_{i=1}^{n} X_i = X_1 X_2 \cdots X_n$$

Show that Y_n converges in probability to a limit Y, possibly a non-random constant, as $n \to \infty$. You need to (1) identify the limit Y, possibly a non-random constant, and (2) prove that Y_n converges in probability to the limit you have identified.

5. Let $X_1, \ldots, X_{10}$ be i.i.d. Binomial(35, p). Suppose we wish to estimate $\theta = 35p(1-p)^{34}$. Let $T = \sum_{i=1}^{10} X_i$. Consider the following estimators of θ.

A: $\mathrm{I}\{X_1 = 1\}$ $\quad$ B: $\frac{1}{10}\sum_{i=1}^{10} \mathrm{I}\{X_i = 1\}$ $\quad$ C: $35\left(\frac{T}{350}\right)\left(1 - \frac{T}{350}\right)^{34}$ $\quad$ D: $35\frac{\binom{315}{T-1}}{\binom{350}{T}}$

(a) Invent a "real" context for this problem. It doesn't have to be the most reasonable context, so try to keep it somewhat simple. Explain in words in your context what p, the X_i's, θ, $T/350$, and estimator B represent.

For each of the following, choose only one estimator (A, B, C, D) and explain your choice. **Your explanation does not necessarily need to include calculations/derivations. A few clearly worded sentences could be sufficient as long as they contain the most relevant ideas applied in this context.** In particular, process of elimination is a valid strategy, though I'm not necessarily suggesting it. Note: some letters (A, B, C, D) might be used more than once; some not at all.

(b) Which estimator is the MLE of θ? Choose only one of (A, B, C, D) and explain.

(c) Three of these estimators are unbiased estimators of θ. Which estimator is NOT an unbiased estimator of θ? Choose only one of (A, B, C, D) and explain.

(d) Which estimator is the uniformly minimum variance unbiased estimator (UMVUE) of θ? Choose only one of (A, B, C, D) and explain. Note: you can assume that except for the estimator you identified in the previous part, the other three estimators are unbiased estimators of θ.

(e) Which estimator has the largest variance? Choose only one of (A, B, C, D) and explain.

6. Xiomara claims that she can predict which way a coin flip will land. To assess her claim, you secretly flip a fair coin 10 times, and she correctly predicts the result in 9 of the 10 flips.

   Rogelio claims that he can taste the difference between Coke and Pepsi. To assess his claim, you give him a blind taste test of 10 cups of soda, secretly flipping a coin for each cup to determine whether to serve Coke or Pespi. Rogelio correctly identifies the soda in 9 of the 10 cups.

   (a) After performing your assessments, whose claim do you find more convincing?
   - Xiomara's claim is more convincing.
   - Rogelio's claim is more convincing.
   - Both claims are equally convincing.

   Choose one and perform an appropriate analysis to justify your choice. Clearly state any assumptions, and compute and report appropriate quantities in context.

   (b) Suggest another way to analyze part a), and state what conclusion the analysis would yield. Then discuss why you prefer your analysis from part a).

7. Suppose the number of home runs hit per game (by both teams in total) at a particular Major League Baseball park follows a Poisson distribution with parameter θ, independently from game to game. In a sample of 5 games there were a total of 9 home runs hit.

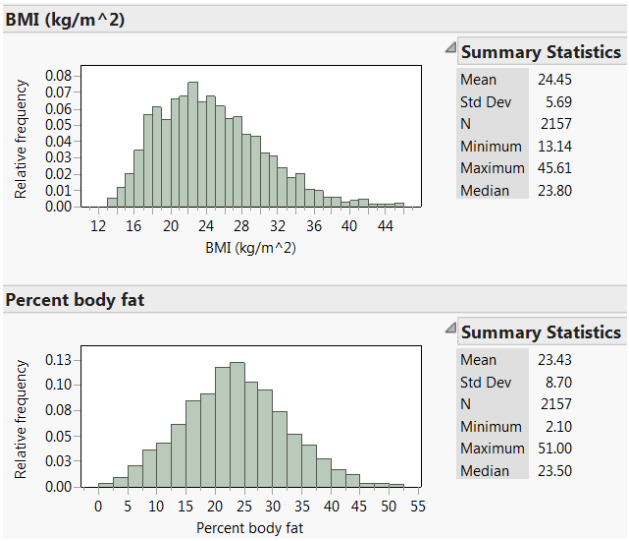
   (a) Assume that the prior distribution of θ is Normal with mean 2 and standard deviation 0.4. Explain how, in principle, you could use simulation to approximate the posterior probability that θ is greater than 3.

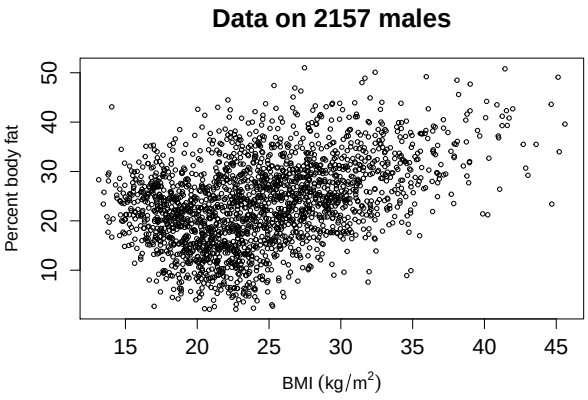
   Note: This question is NOT asking you to describe a sophisticated MCMC algorithm. You are encouraged to use the simplest simulation you can think of, no matter how naive or inefficient it is. This question is really asking you to demonstrate that you know what it means to approximate a posterior probability.

   (b) Now suppose that you have well approximated the posterior distribution, and that you plan to collect data on another sample of 5 games. Explain how you could use simulation to approximate the posterior predictive probability that there will be at least 10 home runs in total in the new sample of 5 games.

8. Let p_{MS} denote the proportion of middle school (MS) students who prefer online to in-person classes, and define p_{HS} and p_C similarly for high school (HS) and college (C) students. You plan to select independent random samples of 100 middle school students, 100 high school students, and 100 college students, and to use the sample data to perform an appropriate Chi-Square test of the null hypothesis $H_0 : p_{MS} = p_{HS} = p_C$. Describe in detail how you could use simulation to approximate the power of the size 0.01 Chi-Square test if $p_{MS} = 0.3$, $p_{HS} = 0.3$, and $p_C = 0.2$.

9. The following summarizes data on a nationally representative sample of U.S. males from the National Health and Nutrition Examination Survey (NHANES).





The correlation coefficient between BMI and percent body fat is 0.4. You fit a standard simple linear regression model for predicting percent body fat from BMI (kg/m^2).

(a) Compute a 95% *confidence* interval corresponding to a BMI of 15 kg/m^2. Write a clearly worded sentence reporting this interval in context.

(b) Compute a 95% *prediction* interval corresponding to a BMI of 15 kg/m^2. Write a clearly worded sentence reporting this interval in context.

10. You should not respond to the following question during the written exam, but please think about it before the oral exam. Which topics that were not covered (or not covered in detail) on this exam do you wish were covered (or covered in more detail)? Why?