2020 Swarthmore Honors Examination in Statistics

Weiwen Miao
Haverford College

May 2020

Instructions: The exam has five questions. Number the questions clearly on your answer sheets. You must make it clear how you arrived at your answer. Answers without any work may lose credit even if they are correct. (Except problem 3(c))

This is a closed-book three-hour exam. You may not refer to notes or textbooks. You may use a calculator that does not do algebra or calculus. The Normal table should be supplied with this exam.

1. Let Y be a random variable with density function

$$f_Y(y;\lambda) = \frac{1}{2\lambda\sqrt{y}} e^{-\sqrt{y}/\lambda}, \ y > 0.$$

   (a) Let $X = \sqrt{Y}$. Find the density function for X.

   (b) Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from the density $f_Y(y;\lambda)$. Find the MLE for λ, call it $\hat{\lambda}$.

   (c) Use (a) to find $E(\hat{\lambda})$. Is $\hat{\lambda}$ unbiased?

   (d) Find the variance of $\hat{\lambda}$.

   (e) Find the Cramér-Rao lower bound. Is $\hat{\lambda}$ an efficient estimator?

   (f) Is $\hat{\lambda}$ consistent?

   (g) Is $\hat{\lambda}$ sufficient?

2. Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with success parameter p. Let $Y = \sum X_i$. Suppose the prior distribution for p is $Beta(\alpha, \beta)$. The density function for $Beta(\alpha, \beta)$ is: $f(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, 0 < p < 1$, where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$ is the gamma function. The mean of $Beta(\alpha, \beta)$ distribution is $\frac{\alpha}{\alpha+\beta}$.

(a) Find the posterior distribution for p.

(b) Using the square-loss function, find the Bayesian estimator for p. Call it $\tilde{p}$

(c) Find the MSE (Mean-Square-Error) for $\tilde{p}$.

(d) It is well-known that the MLE for p is $\hat{p} = \frac{Y}{n}$. Find the MSE for $\hat{p}$.

(e) Suppose $\alpha = \beta = 2$ and $n = 16$. Which estimator will you choose, based on MSE?

3. The *Waisome v. Port Authority of New York* concerned possible adverse impact on African American candidates in the promotional process of police officers to the rank of sergeant. In order to be promoted, candidates needed to pass a written exam. In *Waisome*, 50 of 64 African American applicants passed the written exam while 455 of 508 whites passed.

   (a) Find the plus-four 95% confidence interval for the difference between the pass rates of African American and white candidates.

   (b) Do you think the data provide strong evidence that the pass rate for African American candidates is significantly different than that of white candidates? State the null and alternative hypothesis, calculate the test statistic and the approximate p-value, then draw a conclusion. Test at 0.05 level of significance.

   (c) Yes or No: Can you use the confidence interval in (a) to make the decision in (b)? (Only the answer will be graded)

   (d) A widely used rule of thumb for deciding whether an employment practice has adverse impact is the government's four-fifths rule, which states that the agencies "will generally consider a selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5ths) or eighty percent (80%) of the selection rate for the group with the highest selection rate as a substantially different rate of selection" (Uniform Guideline Q & A #11). Furthermore, by the so-called "flip-flop rule", "If only one more black had been hired instead of a white the selection rate for blacks would be higher than that for whites", then the "Federal agencies will not assume the existence of adverse impact" (Uniform Guideline Q & A #21).

   In *Waisome*, the district court noticed that the ratio of the two pass rates was 87.2%. Furthermore, the court also used an *analogue* of the "flip-flop" rule: the court pointed out that in practical terms, had two more African Americans (hence two fewer whites) passed the exam, the difference would no longer be statistically significant. Consequently, the district court followed the four-fifths rule and concluded that the written exam had no adverse impact on African American candidates.

   The remaining questions explore the district court's requirement for adverse impact: at the 0.05 level of significance, the pass rate for African Americans must be significantly different than that of whites even if two more African Americans and two fewer whites had passed the exam.

Let X be the number of African Americans who pass the exam, Y be the number of whites who pass the exam. Assume that 64 African Americans and 508 whites took the exam. Let $\hat{p}_1 = X/64$ be the pass rate for African Americans, $\hat{p}_2 = Y/508$ be the pass rate for whites. Furthermore, assume that the test is set such that the overall pass rate is $p = \frac{50+455}{64+508} = 88.3\%$.

When both races have the same pass rate of $p = 88.3\%$, what is the asymptotic distribution of $\hat{p}_1 - \hat{p}_2$?

(e) Does the district's requirement make it easier or harder for plaintiffs to show the existence of adverse impact?

4. It's well known that the IQ scores follow a normal distribution with mean 100 and standard deviation 15. Your statistics professor suspects that the average IQ score for Swarthmore students is higher than the national average of 100. Hypothetically, a random sample of size 36 of Swarthmore students yields an average of 105 (fake data). Suppose the standard deviation of the IQ scores for Swarthmore students is also 15 points.

   (a) Do you think the data provide strong evidence that the average Swarthmore student's IQ score is higher than 100? Test at 0.05 level of significance.

   (b) Find the power of the test if the average IQ score for Swarthmore students is 106.

   (c) If the professor wants the power of the test to be 0.9 when the true average IQ is 106, how many Swarthmore students should the professor sample?

   (d) The professor decides to only sample five students and to reject the null hypothesis if at least one student has IQ score higher than 130. Find the probability of committing type I error using the professor's criterion.

5. Let $(x_1, y_1), \ldots (x_n, y_n)$ be a set of n data points. Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ be the mean of x and y, respectively. Let $\hat{y}_i$ be the least-squares predicted value, $i = 1, 2, \ldots, n$. Prove the following facts about the least-squares regression line.

   (a) The least-squares regression line always passes through the point $(\bar{x}, \bar{y})$.

   (b) $\sum y_i = \sum \hat{y}_i$.

   (c) Let $Var(y)$ and $Var(\hat{y})$ be the variance of y and $\hat{y}$, respectively. Show that $R^2 = \frac{Var(\hat{y})}{Var(y)}$.