Brown University
DATA 1010
Fall 2018: Practice Final
Samuel S. Watson

Name:

You will have three hours to complete the exam, which consists of 40 questions. Among the first 36 questions, you should only solve problems for standards for which you want to improve your medal from the second exam.

No calculators or other materials are allowed, except the provided reference sheets.

You are responsible for explaining your answer to **every** *question. Your explanations do not have to be any longer than necessary to convince the reader that your answer is correct.*

For questions with a final answer box, please write your answer as clearly as possible and strictly in accordance with the format specified in the problem statement. Do not write anything else in the answer box. Your answers will be grouped by Gradescope's AI, so following these instructions will make the grading process much smoother.

I verify that I have read the instructions and will abide by the rules of the exam: ____________________________

Problem 37 **[POINTEST]**

(a) Consider the statistical functional $T(\nu)$ which returns the second moment of ν (in other words, $T(\nu) = \mathbb{E}[X^2]$ where X is ν-distributed), and let $\theta = T(\nu)$. Is the plug-in estimator of θ biased? Is it consistent?

(b) Now consider the estimator $\widehat{\theta}$ of θ which is defined to be the sum of (i) the square of the plug-in estimator of the mean of ν and (ii) the plug-in estimator of the variance of ν. Is $\widehat{\theta}$ biased? Is it consistent?

Solution

(a) The plug-in estimator of θ is $\frac{1}{n}\sum_{i=1}^{n} X_i^2$, which is unbiased by linearity of expectation and consistent by the law of large numbers.

(b) We have

$$\begin{aligned}
\widehat{\theta} &= \overline{X}^2 + \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= \overline{X}^2 + \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\overline{X}\frac{1}{n}\sum_{i=1}^{n} X_i + \overline{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2.
\end{aligned}$$

Therefore, this estimator is actually the same as the estimator in (a), and it is therefore also unbiased and consistent.

Problem 38 **[BOOT]**

One thousand voters are polled about their position on a given ballot initiative, and 637 of them respond that they are in favor of the initiative.

(a) Find the value of the plug-in estimator $\widehat{p}$ of the proportion p of voters who are in favor of the initiative.

(b) Write an expression which approximates the standard error of $\widehat{p}$.

(c) Describe how the bootstrap methodology would be used to produce an estimate of the standard error of $\widehat{p}$. Which approach do you find preferable in this case?

Solution

(a) The value of the plug-in estimator for the observed samples is $\widehat{p} = 637/100 = 63.7\%$.

(b) We can approximate the standard error of $\widehat{p}$ in terms of the value of $\widehat{p}$ using the fact that the variance of a binomial random variable is $np(1-p)$, and therefore the variance of a binomial random variable divided by n is $p(1-p)/n$. So we estimate the standard error as

$$\sqrt{\frac{(0.637)(1-0.637)}{1000}}.$$

(c) We could use the bootstrap to accomplish the same objective by drawing from the 1000 responses 1000 times with replacement, calculating the value of the estimator $\widehat{p}$ for each of them, and computing the sample variance of the resulting list.

In this case, using the formula is preferable, since it provides the exact limiting value of the bootstrap procedure with far less computational expense. The reason that the limiting bootstrap value is equal to the value returned by the formula is that drawing with replacement from 1000 samples, 637 of which are 1's, is exactly the same as sampling independent Bernoulli random variables with $p = 63.7\%$.

Problem 39 **[HYPTEST]**

One Bayesian criticism of the hypothesis test framework is that it doesn't account for the *a priori* plausibility of the alternative hypothesis.

(a) You have a magician's coin, and you don't know whether it's a regular coin or a two-headed or two-tailed coin. Consider the null hypothesis that the coin is fair, with the alternative hypothesis that the coin favors one of the two sides. You flip the coin 10 times, and it comes up heads all 10 times. The null hypothesis is rejected with what p-value? What do you actually believe about the coin?

(b) Now suppose you have a coin that you just got from the cashier at Trader Joe's. You carefully inspect it and determine that it appears to be an entirely ordinary U.S. quarter. Once again, consider the null hypothesis that the coin is fair, with the alternative hypothesis that the coin favors one of the two sides. Once again, you flip the coin 10 times, and it comes up heads all 10 times. What do you actually believe about this coin?

Solution

(a) Under the null hypothesis, the probability of getting all 10 heads or all 10 tails is $2/1024 = 1/512$. Therefore, the p-value is $1/512$. I would believe that this coin is likely two-headed.

(b) The p-value is the same as in (a), but I would believe that the coin is fair and it just happened to come up heads 10 times in a row.

Problem 40 **[MLE]**

(a) Find the maximum likelihood estimator for the family of geometric distributions with parameter $0 < p < 1$. (You don't need to prove that the value you find is actually a maximum; just differentiate the log-likelihood and solve for the zero).

(b) I simulated 10 independent samples from a geometric distribution with parameter p and got

$$0, 4, 1, 3, 4, 3, 1, 14, 0, 13$$

Use the maximum likelihood estimator to estimate the value of p that I used.

Solution

(a) Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables with common distribution Geometric(p). The likelihood function is

$$\mathcal{L}(p) = p(1-p)^{X_1 - 1} p(1-p)^{X_2 - 1} \cdots p(1-p)^{X_n - 1},$$

so the log likelihood is

$$n \log p + \left(\sum_{i=1}^{n} X_i - n \right) \log(1-p).$$

Setting the derivative of the log likelihood equal to zero yields

$$\frac{n}{p} - \frac{\sum_{i=1}^{n} X_i - n}{1-p} = 0,$$

and solving for p gives $p = n / \sum_{i=1}^{n} X_i = 1/\overline{X}_n$.

(b) The average of the 10 provided values is 4.3, so the maximum likelihood estimator of p is $1/4.3$.