Econ 21020 - Problem Set 2

Due 10/20 at 11:59 PM. Submit to Canvas. May be completed in groups of up to 6 students. Only one submission is required per group.

Problem 1

Complete the proof of the biasedness of the sample variance that we began in class. That is, demonstrate that, for $X_1, ..., X_n$ iid $\sim X$

$$E[(\overline{X}_n - E[X])^2] = \frac{1}{n} Var(X)$$

Justify all steps. (Hint: We learned that for Y mean independent of X, $E[YX] = E[Y]E[X]$. We also learned that independence implies mean independence. Thus, the former property also holds for independent X and Y).

Problem 2

Suppose that Y_a and Y_b are Bernoulli random variables where $Y_a \sim$ Bernoulli(p_a) and $Y_b \sim$ Bernoulli(p_b). An iid sample of size n_a is drawn from $\sim Y_a$ and another iid sample of size n_b is drawn from $\sim Y_b$. Say that $\hat{p}_a$ is the proportion of the first sample that has a value of 1 and that $\hat{p}_b$ is the proportion of the second sample that has a value of 1. Assume that both samples are also independent of one another.

(a) Show that $\hat{p}_a - \hat{p}_b$ is an unbiased estimator for $p_a - p_b$.

(b) Derive $Var(\hat{p}_a - \hat{p}_b)$

(c) Assume that both n_a and n_b are large. Show what a 95% confidence interval for $p_a - p_b$ would look like in terms of $\hat{p}_a$, $\hat{p}_b$, n_a, n_b, and numbers.

Problem 3

Consider a sample $X_1, ..., X_n$ that is iid $\sim X$.

(a) Suppose that $X \sim N(\mu, \sigma^2)$. What is the sampling distribution for $\overline{X}_{10}$ (the sample mean when we have a sample of 10)?

(b) Suppose instead that $X \sim$ Bernoulli(p). Would the sampling distribution from part a) still apply for $\overline{X}_{10}$? (If not, no need to show what it would be instead).

(c) Find the limiting distribution for $\sqrt{n}(\overline{X}_n - E[X])$ - that is consider the distribution for this expression as n grows arbitrarily large - for both $X \sim N(\mu, \sigma^2)$ and $X \sim$ Bernoulli(p).

Problem 4

Let $(X_1, Y_1), ..., (X_n, Y_n)$ be iid $\sim (X, Y)$. Assume that $E[X^2] < \infty$, $E[Y^2] < \infty$, and $E[(XY)^2] < \infty$. Show that the sample covariance

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}_n \overline{Y}_n$$

is a consistent estimator for $Cov(X, Y)$ using the WLLN and CMT. (Hint 1: This will look quite a bit like the proof of consistency for the sample variance). (Hint 2: Remember that the CMT can be applied with any finite number of sequences of random variables that converge to any finite number of scalars. Alternatively, you can consider applying the CMT twice in a row).

Problem 5

This problem will walk through the development of one-sided hypothesis tests, p-values, and confidence sets that are analogous to the two-sided versions we saw in class. For everything that follows, assume that $X_1, ..., X_n$ are iid $\sim X$ and that $0 < \sigma_X^2 < \infty$.

(a) We have a null hypothesis, $H_0 : E[X] = \mu_0$, alternative hypothesis, $H_1 : E[X] > \mu_0$, and a test statistic:

$$T_n = \frac{\sqrt{n}}{\hat{\sigma}_n}(\overline{X}_n - \mu_0)$$

We want to achieve the significance level α for our test. Determine the critical value c that will allow us to achieve the desired significance level, if our decision rule is to reject H_0 when $T_n > c$. (NB: We could use the same critical value you'll find here if we had the null $H_0 : E[X] \leq \mu_0$ instead).

(b) We are using a (slightly) different test statistic than we saw in class. Explain why in words.

(c) Our decision rule is that we will reject the null if $T_n > c$. Using the given test statistic and the critical value you derived in part a), express the one-sided p-value in terms of numbers and functions that we already know (we

"know" μ_0 as we chose this in advance) and things we can estimate from the data.

(d) Is this one-sided p-value weakly greater or weakly smaller than the two-sided p-value we saw in class? Explain the intuition for the answer.

Problem 6

Complete the following in R or another language of your choice (providing that you have cleared you choice with the TA first).

(a) Generate 100 samples of $n = 5$ from a Uniform[-1,1] distribution. For each sample, compute the sample mean. What proportion of the sample means lie within the range [-0.1,0.1]?

(b) Repeat the above with sample sizes of $n = 10$ and $n = 100$.

(c) Interpret what you find through the lens of results from the class.

(d) For each of the three sample sizes, create histograms plotting 100 values of $\sqrt{n}\overline{X}_n$ (so three histograms - one each showing 100 values of $\sqrt{5}\overline{X}_5$, $\sqrt{10}\overline{X}_{10}$, and $\sqrt{100}\overline{X}_{100}$). Interpret what you find through the lens of results from the class.