Problem Set 4

EC 303: Empirical Economic Analysis

Due November 8, 2019 at 3:30pm.

1 Theoretical Problems

Problem 1.1: An Estimator of Variance. Typically, when estimating variance, we use an estimator of the form $\hat{\sigma}^2 = c\sum_{i=1}^{n}(X_i - \overline{X})^2$, where $c \in \mathbb{R}$ is a constant that may depend on the sample size n.

a. Show that the expected value of $\hat{\sigma}$ is equivalent to $c \cdot g(n) \cdot \sigma^2$, where $g(n)$ is a linear function of the sample size n.

   - *Hint:* Use the fact that $\sum(X_i - \overline{X})^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$, and that $\mathbb{E}[Y^2] = V(Y) + \mathbb{E}[Y]^2$.

b.** What does the value of c need to be for this estimator to be unbiased?

c.** If instead we use $c' = \frac{1}{n}$, what would the bias of the estimator be? Does this biased estimator tend to over- or under-estimate the true value σ^2? How does the bias change as n increases?

d.** Taking as given that $\mathbb{V}[\hat{\sigma}^2] = c^2\sigma^4[2(n-1)]$, what value of c minimizes the MSE? What does this tell you about your estimator?

e. If S^2 is an unbiased estimator for σ^2, is S an unbiased estimator for σ? Why or why not?

Problem 1.2: Standard Errors. Complete Exercise 7.11 from the textbook.

Problem 1.3: Method of Moments **. a. Recall that a Poisson distribution depends on a single parameter λ, which is both its mean and variance ($\mathbb{E}(X) = \mathbb{V}(X) = \lambda$). Use the method of moments to derive two estimates, $\hat{\lambda}_1$ and $\hat{\lambda}_2$, from a simple random sample $\{x_1, x_2, ..., x_n\}$ that has mean 153 and standard deviation 14. Comment on their difference—do you think it is sizeable?

b. Let $\{Y_1, Y_2, ..., Y_n\}$ denote an i.i.d. sample from a population with a $F(\nu_1, \nu_2)$ distribution where ν_1 and ν_2 are the numerator and denominator degrees of freedom, respectively. It can be shown that the mean only depends on the denominator degrees of freedom:

$$\mathbb{E}(Y_i) = \nu_2/(\nu_2 - 2).$$

Use this fact to derive an estimator of ν_2 using the method of moments.

Problem 1.4: Maximum Likelihood Estimation **. Remember to maximize the *log* of the likelihood function in your estimations!

a. Suppose that $\{X_1, X_2, ..., X_n\}$ is an i.i.d. random sample, where X_i has a Bernoulli distribution, with parameter p. Find a maximum likelihood estimator p^{MLE}.

b. Suppose that $\{X_1, X_2, ..., X_n\}$ is an i.i.d. random sample, where X_i follows a normal distribution, with mean zero and unknown standard deviation σ. Find the maximum likelihood estimator $\hat{\sigma}^{\text{MLE}}$. How does this compare to the typical estimator?

Problem 1.5: More Estimation. Complete Exercises 7.23 and 7.30 from the textbook.

Problem 1.6: A More Flexible Confidence Interval. Complete this textbook exercise (8.8): let $\alpha_1, \alpha_2 \in (0,1)$ with $\alpha = \alpha_1 + \alpha_2$. Then

$$P\left(-z_{\alpha_1} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha_2}\right) = 1 - \alpha$$

a. Use this equation to derive a more general version of the confidence interval for μ (the special case that we're used to dealing with is Equation 8.5 in the textbook).

b. Let $\alpha = 0.05$ (what kind of confidence interval does this mean we're constructing?), and let $\alpha_1 = \alpha/4$ and $\alpha_2 = 3\alpha/4$. Will the resulting interval be narrower or wider than the original interval?

Problem 1.7: Constructing Confidence Intervals **. These problems give you practice constructing various confidence intervals.

a. Complete Exercise 8.5 parts (a)–(c) from the textbook.

b. Complete Exercise 8.26 from the textbook.

Problem 1.8: Familiarizing Ourselves with New Distributions. Complete Exercises 8.30 and 8.44 from the textbook. These are quick exercises to get you comfortable with the t and χ^2 distributions.

Problem 1.9: An Application. Finally, complete Exercises 8.34 and 8.46 from the textbook, to put it all together.

2 Stata Exercises

Problem 2.1: Bootstrapping Standard Errors. This problem will walk you through a bootstrap method to obtain an estimator and standard errors for a difference in means and percentiles.

For this exercise, use the discrim data set from the bcuse package. This data contains information on fast food prices and neighborhood characteristics. We are interested in testing the claim that those in high-minority neighborhoods (in this case proxied by the fraction of blacks in a neighborhood) are subject to higher prices due to discrimination.

a. Make a scatter plot of the psoda and prpblck variables, adding a linear trend over top. What do you see? Interpret the linear trend, and argue whether you think the result looks reasonable given the scatter plot.

b. Create a binary indicator for if a neighborhood has 20% or higher black population. Calculate the mean and median soda price in each type of neighborhood (high minority and not). Compare and interpret the differences across neighborhoods, as well as the difference between mean and median.

c. Now bootstrap the difference in means across both neighborhood types. That is, in each replication, your bootstrapped value should be: mean(psoda|prpblck$\geq$ 0.2)$-$mean(psoda|prpblck$<$ 0.2). What is the bootstrapped mean difference in prices (and its standard error)? Interpret this result.

   – Use 500 replications, and don't forget to set the seed to 1223334444.

d. Now repeat the exercise for the difference in the 10th percentile of prices. For this one, use 1000 replications (as there are generally fewer possible outcomes). Interpret this result as before. Why might we be interested in this percentile rather than the mean?

e. We often say that a difference is "significantly" different from 0 if its confidence interval does not include 0. Construct a 95% confidence interval around both of your estimated differences (using the bootstrapped standard error). Are these significant? Do the differences seem economically meaningful to you? What would you conclude about discrimination in fast food prices?

Problem 2.2: Maximum Likelihood Estimation & Confidence Intervals. This problem walks through using MLE and constructing CIs in Stata. First, we will perform maximum likelihood estimation. This can save the rather tedious calculation of derivatives. For example, consider the Poisson distribution:

$$f(y;\mu) = \frac{\mu^y e^{-\mu}}{y!}$$

Calculating these derivatives would be little fun as factorials become unwieldy quickly.

a. Simulate a data set of 5000 observations from a Poisson distribution with a true parameter $\mu = 5$. Call this variable myvar (for ease of following the instructions later on).

b. For an i.i.d. sample of n observations, write and simplify the log-likelihood function for this distribution (no Stata necessary)

c. Now we will walk through how to estimate this using Stata's MLE program:

   – We will specify a program that calculates the log-likelihood function for any given observation. The function we will use is

   $$\ell(\mu|y_i) = y_i \ln(\mu) - \mu - \ln(y_i!).$$

   How does this compare to your answer to (b)?

   – First, define a program that calculates the log-likelihood function for any observation. This program has the following syntax:

```
program define poisson
    version 1.0
    args lnf mu
    quietly replace `lnf' = $ML_y1*ln(`mu')- `mu' - lnfact($ML_y1)
end
```

   The program starts by us assigning a name (poisson) and a version (1.0). We then give it a list of arguments that we care about—the first (lnf) is the log-likelihood that we wish to calculate, and the second (μ) is the parameter to estimate. The last line of the program uses μ and any observation ($ML_y1) to calculate the log-likelihood for that particular observation. Stata will automatically do the summation at the end of the procedure.

   – Once the program has been defined, we call it into action with the command ml model lf poisson (myvar=). This command calls a maximum likelihood model (ml model) that performs a linear fit (lf) using our program and dependent variable myvar.

   – Finally, the estimation is performed with the command ml maximize

   – What is the estimated parameter μ? What is its 95% confidence interval, and does that interval contain the true μ?

d. Would the MLE estimation perform this nicely if the data were a little noisier? To each observation, add a shock $\epsilon \sim \mathcal{N}(0,1)$. Since we are working with a discrete log-likelihood function (we have factorials), you will need to round each observation of the new variable to the closest integer, and make sure that any negative integers are changed to 0 (so that we can take logarithms[1]). Now repeat the estimation in part (b) and report your new estimated $\hat{\mu}$ and its 95% confidence interval. Did the estimation perform worse? Why do you think that is?

e. While Stata gives you the 95% confidence interval easily, it can also quickly calculate any interval you'd like to know. Use the invnorm() function to construct a 92%, 97%, and 99.6% confidence interval for your answer to (b) and (c). Remember that the area to the left of a confidence interval should be $\alpha/2$, not just α. Present your results in a well-organized table (you don't need to make this table in Stata unless you really want to). How do the confidence intervals compare across the two estimation techniques?

[1] Note that Stata can handle ln(0) by using a very large negative number.

Problem 2.3: Bootstrapping Confidence Intervals. In some cases, we may care about finding a confidence interval for a parameter even when n is not large and the underlying distribution is non-normal. Once again, bootstrapping can help us! For this problem, use the prminwge data, which contains a small number of observations on Puerto Rico's minimum and average wages.

a. Report a normality plot of the avgwage variable. Does it appear to come from a normal distribution?

b. Now bootstrap the sample and store the average wage each time. Report a histogram of these average wages. Use 999 replications (you'll see why in a second). Does this appear normal?

   - Don't forget to set your seed.

c. Now we can form a confidence interval using the sample standard deviation of the bootstrap means, s_{boot}:

$$(\overline{X} - z_{\alpha/2} \cdot s_{\text{boot}}, \overline{X} + z_{\alpha/2} \cdot s_{\text{boot}})$$

   What is the 95% confidence interval for the average wage using your bootstrapped estimates?

d. If the bootstrapped distribution is not perfectly normal, we can use a **percentile interval** instead. This uses the $\alpha/2$ and $(1-\alpha)/2$ percentiles of the bootstrapped definition to construct an interval. That is, if we sort the bootstrapped means from smallest to largest, we would choose the k-th smallest and k-th largest estimates, where $k = \alpha(B+1)/2$.

   - This is why we used $B = 1000$. In this case, what is k for a 95% percentile interval?
   - What is the 95% percentile interval given our bootstrapped data? How similar/different is it to the 95% confidence interval? Why do you think that is?

e. Under some circumstances, the actual confidence level may differ substantially from the nominal level (the level you think you are getting); that is, we may not be actually obtaining a 95% confidence interval using the percentile interval approach. To correct for this, Stata can construct a bias corrected and accelerated (BCa) interval. To implement this, re-load the prminwge data set and use the following two lines:

   - bootstrap, reps(999) bca: mean avgwage (performs the actual bootstrapping)
   - estat bootstrap (gives you the corrected confidence interval)

   What is the resulting interval? In what direction did the interval move? What does this imply about the bias of the estimator in this context (e.g., in the context of the average wage of PR)?