Assignment 2

*Note:* Unless otherwise specified, data sets comes from Wooldridge’s econometrics textbook. To use them, use the R package “wooldridge”.

**Problem 1: Working with Normal Distributions.** R can be quite useful in calculating probabilities from a normal distribution. To see this, we will use the openness data set, which contains data on openness to foreign trade for various countries.

1. Construct a nice figure depicting the distribution for open. Does this variable appear normally distributed? What about lopen? What does this imply about the approximate distribution of open?
2. What are the mean and standard deviation of lopen?
3. Use [this website](https://ourworldindata.org/grapher/trade-openness) to find current data on a country’s openness to trade (the link should have the data show immediately; if not, select the variable “Openness at constant prices” from the drop down menu). Select a country of interest and report their (i) openness score (in percentages, not decimals) and (ii) log openness score. Then calculate (iii) the *z*-score using the measures you calculated in part (b). How many standard deviations above/below the mean is the country you selected?
4. Back in R, transform lopen into a new variable that is its *z*-score and verify that the new variable is approximately standard normal. What proportion of observations in the data are below the *z*-score you reported in part (c)?
5. Now report the *exact* probability that a randomly selected country has (i) an openness score lower than the country you chose, and (ii) an openness score between your country and the average openness score. How does your answer to (i) compare to your answer to part (d) above? Why are the numbers different?

**Problem 2: Covariance and Correlation.** For this problem, use the Excel file called “Uninsured.xslx” in the “Datasets” folder on GitHub. This data set contains information on 20 municipalities in Massachusetts. For each municipality, the fraction of people without health insurance (frac uninsured) and the fraction of people declaring bankruptcy (frac bankrupt) are reported.

* 1. What is the covariance between these two variables? Make a nice-looking scatterplot of the variables’ relationship. How does the covariance you reported jibe with the graph? Why do you think this is?
  2. Create new variables for both bankruptcy and (un-)insurance that is measured in people (rather than percentages). Use the population variable to do so. Does this change the linear relationship? What is the new covariance? What does this teach you about covariance and data viz?
  3. As discussed in class, the correlation is a unitless measure that resolves some of the problems discussed above. What is the correlation between the two original variables? Does this correlation change when you use the new variables (based on people, not percentages) instead? Why (or why not)?
  4. What is the correlation between frac uninsured and your count of uninsured? What explains this? Why doesn’t this cause a problem in calculating the correlation between your new variables? (A scatter plot—or multiple plots—may be useful here.)
  5. However, even looking at the correlation can be misleading. Create a data set of 1000 observations and 2 variables: *X* that ranges continuously over the interval [0*,* 5] and *Y* given by *y* = *x*(*x* 5). Create a scatter plot of this relationship. What economic variables may have this relationship? What is the correlation between *X* and *Y* , and what drives this result? When should I be careful of looking only at the correlation coefficient *ρ*?

− −

**Problem 3: 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, which 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.

1. 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.
2. 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.
3. Now bootstrap the difference in means across both neighborhood types. That is, in each replication, your bootstrapped value should be: mean(psoda prpblck 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 for replicability.

1. 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?
2. 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 4: 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.

1. Report a normality plot of the avgwage variable. Does it appear to come from a normal distribution?
2. 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.
3. Now we can form a confidence interval using the sample standard deviation of the bootstrap means,

*s*boot:

(*X* − *zα/*2 · *s*boot*, X* + *zα/*2 · *s*boot)

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

1. If the bootstrapped distribution is not perfectly normal, we can use a **percentile interval** instead. This uses the *α/*2 and (1 *α*)*/*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* = *α*(*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?

**Advanced Problem (for extra practice)**

**Problem 5: Simulation.** This problem asks you to perform simulations as in Chapter 6.

1. 6.10 (revised): Carry out a simulation experiment to study the sampling distribution of *X* when the population distribution is log-normal with E[*ln*(*X*)] = 3 and V[*ln*(*X*)] = 1. Consider the four sample sizes *n* ∈ {10*,* 20*,* 30*,* 50}, and in each case use 500 replications.
   * For each of the 4 simulations, provide a histogram of the sampling distribution.
   * Provide 1 table comparing the means, medians, and standard deviations of the sampling distri- butions across the 4 simulations. Are the sampling distributions symmetric? Are the results consistent with the LLN? Discuss.
   * For which of these sample sizes does the *X* sampling distribution appear to be approximately normal?
   * *Hint:* You will need to use the command set matsize 500 at the beginning of the code for this problem in order to use *k* = 500 in your MATA code.
2. 6.26 (revised): A friend commutes by bus *to and from* work 5 days/week. Suppose that waiting time is uniformly distributed between 0 and 10 min, and that all waiting times are independent of each other.
   * What is the approximate probability that total waiting time for an entire week is at most 60 min? Use *k* = 500 replications.
   * The idea of this problem is that even for an *n* as small as 10, *T*0 and *X* should be approximately normal when the parent distribution is uniform. What do you think?