Homework Set 1, ECO 311, Fall 2014

Due Date: At the beginning of class on September 23, 2014

Instruction: There are ten questions. Each question is worth 2 points. You need to submit the answers of only five questions which you choose. The maximum point you can get is 10 points. I will grade only the first five questions if you answer more than five questions. For the purpose of preparing exam, you need to understand all ten questions.

I will discuss the homework on the due date. Please do not ask me to go through the homework before the due date. However, you can discuss the homework with your classmates. You need to submit the homework individually though.

A warning about notation. The textbook uses uppercase letter such as $Y$ to denote a random variable, and lowercase letter $y$ for the value taken by the variable. Here I use the lowercase letter to denote both to avoid the cumbersome notation.

Q1: Summation

Statistics and Econometrics involve a lot of summations, and the shorthand for summation is $\Sigma$. By definition

$$\sum_{i=1}^{n} y_i \equiv y_1 + y_2 + \ldots + y_n$$

where $i$ is the index for the observation, and $n$ is the number of observations (sample size). For example, we can use the sigma notation to define the sample mean as

$$\bar{y} \equiv \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

More discussion about summation is in appendix A.1 of the textbook. Now please prove

1. $\sum_{i=1}^{n} cy_i = c \sum_{i=1}^{n} y_i$ where $c$ is a constant (1 point)

2. $\sum_{i=1}^{n} (y_i - \bar{y}) = 0$ (0.5 point)

3. $\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y})y_i$ (0.5 point)

The second problem indicates that the sample mean measures the central location, in the sense that the total deviation of $y$ from the sample mean is zero. The third problem shows how to simplify the calculation of sample variance.
Q2: Expectation and Variance

Statistics uses random variable to describe randomness. A variable is random if it follows a distribution (taking different values with probability). The central location of the distribution is called expected value or mean, denoted by $E(y)$ or $\mu_y$. Mathematically, the mean is a weighted average of the values that can be taken by the random variable; the weight is the probability that the variable takes that value (Read appendix B.3 of the textbook):

$$E(y) = \sum_{j=1}^{k} y_j f(y_j)$$

where $j$ is the index for the value that can be taken by $y$. The above formula assumes the variable can take $k$ possible values; the probability for taking the $j$-th value is $f(y_j)$.

Now consider a Bernoulli random variable that takes values of 1 and 0 only (later we call such variable a dummy variable)

$$y = \begin{cases} 
1, & \text{with probability } p \\
0, & \text{with probability } 1 - p 
\end{cases}$$

so $f(1) = p, f(0) = 1 - p$.

1. Prove that $E(y) = p$ (1 point). We learn from this problem that for Bernoulli random variable the parameter $p$ denotes both the mean and the probability that $y = 1$.

2. The dispersion of the distribution is measured by variance:

$$\var(y) = E\left(\left[y - E(y)\right]^2\right).$$

Please find the variance for the Bernoulli random variable $\var(y) =$? (1 point)

Hint: for the second part we may define a new random variable $z = [y - E(y)]^2$. So $z$ is a function of $y$, and $z$ measures the squared deviation of $y$ from its mean. Now you may first find the distribution for $z$, and then the mean of $z$, which is $\var(y)$ by definition.
**Q3: Covariance**

Read Appendix B.4 of the textbook for this problem. The expectation has the properties that for a constant \( c \)

\[
E(c) = c \\
E(cy) = cE(y)
\]

Suppose there are two random variables \( x \) and \( y \). We can define another one

\[
z = (x - \mu_x)(y - \mu_y)
\]

where \( \mu \) is the expected value. Then by definition, the covariance between \( x \) and \( y \) is just the mean of \( z \) :

\[
\text{cov}(x, y) = E(z) = E[(x - \mu_x)(y - \mu_y)]
\]

Use the property of the expectation to prove that

1. \( \text{cov}(x, y) = E(xy) - \mu_x\mu_y \) (1 point). (Hint: \( \mu_x \) is a constant, so is \( \mu_y \))

2. Under what condition it follows that \( E(xy) = E(x)E(y) \)? (1 point)

The last problem shows that in general the expectation of a product is not the product of expectations.

**Q4: Measurement Error and Bias**

This problem intends to make the idea more explicit that inaccurate observations (or observations with error) can lead to inaccurate (biased) estimate. Suppose we are interested in estimating the population mean \( \mu_y = E(y) \) using the random (i.i.d) sample \((y_1, y_2, \ldots, y_n)\). The observed value \( y_i^* \), however, equals the true value plus an error \( e_i \) (called measurement error)

\[
y_i^* = y_i + e_i
\]

1. Find the expression for the sample mean using the inaccurate sample \( \frac{\sum_{i=1}^{n} y_i^*}{n} \) = ? (0.5 point)

2. Find the expected value of the sample mean using the inaccurate sample \( E\left(\frac{\sum_{i=1}^{n} y_i^*}{n}\right) = ? \) (1 point)
3. Under what condition \( \frac{\sum_{i=1}^{n} y_i}{n} \) can still be an unbiased estimate for \( \mu_y \)? (0.5 point)

(Hint: read Appendix C.2 of the textbook)

Q5: Normal Distribution and Law of Large Number

Suppose \( y_i, i = 1, 2, \ldots, n \) are i.i.d random variables (so the sample is random sample), and each is normally distributed \( N(10, 4) \). Denote the sample mean by \( \bar{y} \).

1. What is the distribution of \( \bar{y} \)? Use Table G.1 in the textbook to find \( P(9.6 \leq \bar{y} \leq 10.4) \) when \( n = 25 \).

2. Suppose \( c \) is a positive number. Show that \( P(10 - c \leq \bar{y} \leq 10 + c) \) becomes close to 1 as \( n \) grows large. This result implies that the sampling distribution of the sample mean \( \bar{y} \) becomes more and more concentrated around the population mean as sample size gets larger. This fact is called law of large number.

Q6: Stata Exercise: Correlation Coefficient and Outlier

Use Our Data posted in my webpage to answer this question:

Please use stata to find the variance, covariance of gpa201 and gpahs, and correlation coefficient between them. Please print out the stata result (you may copy and paste it to Word). Interpret the correlation coefficient you get. Then do the analysis again by excluding the student with gpahs=9 (the outlier). This problem illustrates that sometimes it makes a big difference by excluding the outlier.

Q7: Stata Exercise: T test and Confidence Interval

Use Our Data posted in my webpage to answer this question:

First run the command \texttt{sum gpa201}. Based on the result, please find standard error and t statistic for the null hypothesis \( H_0 : \text{mean}(gpa201) = 3.2 \). Use the stata function \texttt{ttail} to find the \( p \)-value and draw a conclusion. Finally please use the stata function \texttt{invttail} to find the critical values and construct the 90% confidence interval for the population mean of gpa201, and interpret that interval. Show me the solving process, i.e., write down the math formula, then plug in the number. Do not use the command \texttt{ttest} to obtain the answer.
Q8: Stata Exercise: Two Sample T Test

Some of my colleagues argue that Miami should have more female students simply because they are doing better in class than boys. Please use Our Data posted in my webpage to run a Two-sample T test to check this issue. To fix idea, please report the two-sample t test that compares the mean of gpa202 across boys and girls. Is this one-tailed or two-tailed test? Why? How much is the p-value? What is your conclusion?

Q9: Sample Mean is OLS Estimator

Consider running a strange regression and only estimating an intercept (i.e., there is no independent variable $X$ on the right hand side of the regression). The regression is

$$y = \beta_0 + u.$$

1. Please show that the estimated intercept is the sample mean

$$\hat{\beta}_0 = \bar{y}$$

2. Please show the sum of residual equals zero

Q10: Stata Exercise: Simple Regression

Use Our Data posted in my webpage to answer this question:

We are interested in the relation between gpa201 and gpahs. The stata command

```
reg gpa201 gpahs
```

reports the below estimated regression

$$\hat{gpa201} = 2.199 + .298gpahs$$

1. Is it likely that gpahs is exogenous (so is uncorrelated with the error term)

2. How to interpret the estimated slope $\hat{\beta}_1 = .298$? Does gpahs have significant effect on gpa201? Why?