Homework 1.5

Due Monday July 16, 2012 *(before class starts!)*

This assignment is designed to help you reinforce the Central Limit Theorem and sampling distribution. They are tedious, but your instructor hopes that doing them will help you in the long run. The first two questions deal with the case when the population is assumed to be Normally distributed. The third question is an exercise in CLT. The fourth question ties CLT and the techniques developed in question one and two.

1. Assume $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $i = 1, ..., n$ for the following questions. Provide mathematical justification to all the questions below.

   (a) Show that the sample mean, $\frac{1}{n}\sum_{i=1}^{n} X_i$, has the distribution $N(\mu, \frac{\sigma^2}{n})$.

   (b) Suppose we transform all of our sample from $X_i \to Y_i$ by the following formula $Y_i = aX_i + b$ where a and b are some constant numbers. What is the distribution of Y_i? Are Y_i's independent of each other? Are they identically distributed?

   (c) Show that the sample variance of X_i is identical to sample variance of Y_i if $a = 1$. Mathematically speaking, show

   $$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

   if $a = 1$. What does this imply about estimates of the sample variance under different transformations.

   (d) Now suppose a can be any arbitrary number. How much does the sample variance of X_i differ in comparison to the sample variance of Y_i? In particular, what is the ratio of the sample variances $\frac{\hat{\sigma}^2_X}{\hat{\sigma}^2_Y}$ where $\hat{\sigma}^2_X$ is the sample variance for X and $\hat{\sigma}^2_Y$ is the sample variance for Y

   (e) What is the sample correlation between $(X_1, ..., X_n)$ and $(Y_1, ..., Y_n)$?

2. Assume $X_i \overset{\text{iid}}{\sim} N(\mu_X, \sigma^2)$, $i = 1, ..., n$ and $Y_i \overset{\text{iid}}{\sim} N(\mu_Y, \sigma^2)$. Provide mathematical justification to all the questions below.

   (a) What parameter is the following statistic

   $$\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} Y_i$$

   trying to estimate? More specifically, what parameter is this estimator unbiased for?

(b) Assume that X_i and Y_i are independent of each other. That is, all X_i's are independent of all Y_i's. Further assume that either μ_X is zero or μ_Y is zero, but you don't know which μ is zero. Compute the bias and find the sampling distribution of the following estimators of σ^2.

$$\text{Estimator 1: } \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$\text{Estimator 2: } \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2+(Y_i-\bar{Y})^2$$

$$\text{Estimator 3: } \frac{1}{2n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2+(Y_i-\bar{Y})^2$$

$$\text{Estimator 4: } \frac{1}{2n-2}\sum_{i=1}^{n}(X_i-\bar{X})^2+(Y_i-\bar{Y})^2$$

$$\text{Estimator 5: } \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}+Y_i-\bar{Y})^2$$

Note that for the sampling distribution, you may have to multiply the estimators by certain constants.

(c) Again, assume that X_i and Y_i are independent of each other. Assume that you know the population mean of μ_X. How does the estimator

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2+(Y_i-\bar{Y})^2$$

compare to the estimator

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu_X)^2+(Y_i-\bar{Y})^2$$

and

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_X)^2+\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

in terms of bias for σ^2?

(d) Suppose we know σ^2. We want to estimate μ_X by using both X_i and Y_i. We know that μ_X is related to μ_Y by the following

$$\mu_X=\mu_Y+a$$

where a is unknown.

(e) Come up with an statistic/estimator for a. Show that this estimator is unbiased for a. What is its sampling distribution?

(f) Let T be the estimator you constructed from the previous question. Consider the following estimators for μ_X

$$\text{Estimator 1: } \frac{1}{n}\sum_{i=1}^{n}X_i$$

$$\text{Estimator 2: } \frac{1}{n}\sum_{i=1}^{n}X_i+\frac{1}{n}\sum_{i=1}^{n}(Y_i-T)$$

Are both of them unbiased? Which statistic has the lowest risk? What are the sampling distributions for each?

3. Suppose $X_i \overset{\text{iid}}{\sim} F$ where F is some arbitrary distribution with mean μ and variance σ^2. We are going to use the following version of the CLT. This version is THE version you can use for all questions in future assignments and quizzes.

Theorem 1. *(CLT) Let $X_i \overset{\text{iid}}{\sim} F$ where F is some arbitrary distribution with mean μ and variance σ^2. Then the distribution*

$$\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} \to N(0,1)$$

as $n \to \infty$.

For all of these questions, provide mathematical justification by directly using the above theorem.

(a) Show that $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$ as n goes to infinity by using theorem 1.

If you showed this correctly, you can show that $\bar{X}$ can be approximated by $N(\mu, \frac{\sigma^2}{n})$, for large n, with the following (and proper) mathematical justification.

Consider the cdf of $\bar{X}$, $P(\bar{X} \leq x)$ where x is any number. Now, based on the fact that $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$, we can say that $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0,1)$ for large sample size. Then,

$$P(\bar{X} \leq x) = P(\bar{X}-\mu \leq x-\mu) = P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}}\right) \approx P\left(Z \leq \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = \Phi\left(\frac{x-\mu}{\frac{\sigma}{\sqrt{n}}}\right) \tag{1}$$

where $\Phi()$ is the cumulative distribution function of the standard normal[1] Now, consider the random variable $Y \sim N(\mu', \sigma'^2)$. To find the probability $P(Y \leq x)$ where x is some number, we always used the z-score trick

$$P(Y \leq x) = P\left(\frac{Y-\mu'}{\sigma'} \leq \frac{x-\mu'}{\sigma'}\right) = P\left(Z \leq \frac{x-\mu'}{\sigma'}\right) = \Phi\left(\frac{x-\mu'}{\sigma'}\right) \tag{2}$$

Thus, if you can write your probability expression in terms of $\Phi()$, you can convert it back to $Y \sim N(\mu', \sigma^2)$ notation. For our approximation for $\bar{X}$, we can combine 1 and 2 to obtain the following result.

$$P(\bar{X} \leq x) \approx \Phi\left(\frac{x-\mu}{\frac{\sigma}{\sqrt{n}}}\right) \Rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

This *should* be the mathematical justification you should provide in your assignments and quizzes. The following questions will attempt to give you practice in working with this type of mathematical justification.

[1]Mathematically, for any z, $P(Z \leq z) = \Phi(z)$. For example, if $z = 1$, then $P(Z \leq 1)$ is the probability that the standard normal is less than one, which is denoted as $\Phi(1)$.

(b) Assume that $\kappa^2 = Var((X_i - \mu)^2)$. Show that

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2}{\frac{\kappa}{\sqrt{n}}} \to N(0,1)$$

by directly using theorem 1. Also, show that the approximate distribution for $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ is $N(\sigma^2, \frac{\kappa^2}{n})$ by following the mathematical justification similar to question 3(a)

(c) Suppose $X_i = 1$ with probability p and -1 otherwise. Show that

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - (2p-1)}{\frac{2\sqrt{p(1-p)}}{\sqrt{(n)}}} \to N(0,1)$$

What is the approximate distribution for $\frac{1}{n}\sum_{i=1}^{n} X_i$? Provide mathematical justification similar to question 3(a).

(d) Suppose $X_i = 1$ with probability p and 0 otherwise. Show that

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} \to N(0,1)$$

This is called the Normal approximation to the Binomial distribution. What is the approximate distribution for $\sum_{i=1}^{n} X_i$? Provide mathematical justification similar to question 3(a).
Hint: You must be careful here with the continuity correction. Remember, if x is a whole number between $0, 1, 2, ..., n$, $P(\sum_{i=1}^{n} X_i \leq x) = P(\sum_{i=1}^{n} X_i \leq x + 0.5)$

(e) Suppose $X_i \sim Exp(\lambda)$. Come up with the limiting distribution for $\frac{1}{n}\sum_{i=1}^{n} X_i$ so that as n goes to infinity, this limiting distribution converges to the standard normal. What is the approximate distribution for $\frac{1}{n}\sum_{i=1}^{n} X_i$? Provide mathematical justification similar to question 3(a).

(f) Suppose $X_i \sim Unif(0,1)$ is a uniform random variable between 0 and 1. Derive the expression for the limiting distribution of $\frac{1}{n}\sum_{i=1}^{n} X_i$ and its approximate distribution. Provide mathematical justification similar to question 3(a).

(g) Suppose $X_i \sim Unif(-1,1)$ is a uniform random variable between -1 and 1. Derive the expression for the limiting distribution of $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ and its approximate distribution. Provide mathematical justification similar to question 3(a).

4. Consider the cell count example we seen in class. To count the total number of cells in the circular microscope's field view with radius r, we decided to consider a small circle, of radius r_s, whose center is chosen randomly on the field view, and count the cells in the smaller circle and multiply by the ratio of the area of the microscope's field view and the area of the small circle. Suppose $\hat{N}_s$ is the number of cells you counted in the small circle and N is the total number of cells in the circular microscope's view. Assume that N cells are distributed uniformly in the microscope's field view

(a) Write down the estimator for the total number of cells in the microscope's field view. Call this estimate T.

(b) What is the expected number of cells in the small circle
Hint: You should use geometry to get an intuitive idea as to what the expected number should look like. However, to argue this rigorously, consider counting the cells in the following manner. Let $X_i = 1$ *if the* i*th cell is in the small circle and* 0 *otherwise. Then* $\sum_{i=1}^{n} X_i$ *is the total number of cells in the small circle. Can you guess what the expectation of this sum of random variables is? Remember, you still have to find* $P(X_i = 1)$*!*

(c) Based on your previous result, what is the bias of T in estimating N, the total number of cells in the field view? Is T an unbiased estimator?

(d) What is the sampling distribution of $\hat{N}_s$, the number of cells you counted in the small circle?
Hint: Use the $\sum_{i=1}^{n} X_i$ *expression in 4(b)'s hint*

(e) What is the expression for the limiting distribution of T? What is the approximate distribution for T, with the continuity correction (see 3(d))? Provide mathematical justification similar to question 3(a)

(f) Does the theory match up with the empirical results we see in the simulation? Are there any bias in the simulation? Where is this bias coming from?