Homework 3

This HW reviews the normal distribution, confidence intervals and the central limit theorem.

0) Go to the link [http://onlinestatbook.com/stat_sim/sampling_dist/index.html](http://onlinestatbook.com/stat_sim/sampling_dist/index.html) (if you are looking at this file from a browser you can simply click on the link).

   (a) By dragging your mouse over the parent population plot, plot a parent distribution of your choice (with several modes etc).

   (b) Think of a population which matches this distribution shape.

   (c) To see how a typical sample behaves, press animate (the size of the sample drawn depends on how you choose N, the choices are N=2,5,10,16,25).

      On the below plot you will see the average of this sample.

   (d) Do this same routine 10000 times (you don’t have to do it yourself!, just press the 10000 button!).

   (e) Do the above for N = 2, 5, 10, 16 and 25 and explain what happens to the distribution (histogram) of the sample mean (both the shape and spread) as N increases.

1) Suppose that X is a normally distributed random variable where $X \sim N(75, 3^2)$ (mean 75 and standard deviation 3).

   (i) Calculate $P(X > 67)$.

   (ii) Find the $x$ such that $P(75 - x \leq X \leq 75 + x) = 0.99$.

2) A patient is classified as having gestational diabetes if the glucose level is above 140 miligrams per deciliter one hour after ingesting a sugary drink. Lucy’s measured sugar level varies according to a normal distribution with mean $\mu = 125 mg/dl$ and standard deviation $10 mg/dl$.

Since the her mean level is below 140 mg/dl she does not have gestational diabetes. However, in reality the mean level is unknown, all that is known are readings taken from blood samples. Therefore, below we want to evaluate the chance of wrongly diagnosing gestational diabetes based on the samples taken.

   (a) Suppose one single measurement is made (one blood sample), what is the probability that she will be misdiagnosed as having gestational diabetes (in other words what is the chance that her measurement will be above 140 mg/dl given that a single measurement is normally distributed with mean $\mu = 125 mg/dl$ and standard deviation $10 mg/dl$)?

   (b) Instead suppose that on three separate days measurements are made and the average measurement is taken over these three days. What is the probability that she will be misdiagnosed as having gestational diabetes (in other words what is the chance that her average over these three measurements will be above 140 mg/dl)?

   *Hint: What is the distribution of the sample mean based on three measurements given that a single measurement is normally distributed with mean $\mu = 125 mg/dl$ and standard deviation $10 mg/dl$?*
(c) Compare your solutions from part (a) and part (b). What have you notice about the probability of false diagnosis as a larger sample is used?

(d) Suppose the distribution of Jane’s glucose samples follow a normal distribution with mean level of 139 mg/dl and standard deviation of 10 mg/dl. As this is not above 140 Jane is not diabetic. However, predict the probability of her being falsely diagnosed based on just one blood sample.

(3) Suppose the scores of high school ACT test have mean 19.2 and standard deviation 5.1. As we discussed in class, ACT scores are only very approximately normally distributed.

(a) Using the normal distribution, what is the approximate probability that a single randomly selected student will score 23 or higher?

(b) A simple random sample of 25 students is taken. What is the mean and standard deviation of the average score (sample mean $\bar{x}$) of these 25 students?

(c) Using the normal distribution, what is the approximate probability that the sample mean score of these 25 randomly selected students will be 23 or higher?

(d) Which of your Normal probability calculations (a) and (c) will be the most accurate, give a reason for your answer?

(4) (i) 300 different samples are drawn, each sample is of size 50 ($n = 50$) (this is a little like the applet you played with in Q1, where several samples are drawn and this distribution of the sample mean is plotted). For each sample, a 90% confidence interval (CI) for the mean $\mu$ is constructed. On average, how many of these 300 confidence intervals will contain the mean?

(ii) Suppose it is known that the smallest adult is 1.5 feet tall and the tallest known adult is 8.5 tall. A sample of size 50 people is drawn, the average height using this sample is 5.5 feet tall.

Give a 100% CI for the mean adult height.

(iii) Suppose a random sample of size 40 is drawn from a population which has mean $\mu$ and variance $\sigma^2$. I evaluate the sample mean $\bar{X} = \frac{1}{40} \sum_{i=1}^{40} X_i$. It is known the standard error of the sample mean is 0.5. What is the standard deviation of the original population?

(5) A random sample of size 15 is drawn. The QQplot is given below. Suppose that the sample mean is $\bar{X} = 0.606$ and the population variance is $\sigma^2 = 1$.

<a href="image.png" download="image.png">QQplot.png</a>

(a) Construct a 95% CI for the mean.

(b) Based on the QQplot comment on whether the 95% CI for the mean is reliable. Give a reason for your answer.
(6) Suppose that the population has an unknown mean ($\mu$) and variance 10 respectively. The distribution of this population is bimodal (has two major peaks).

A random sample of size 30 is drawn from this population and evaluate the sample mean $\bar{X} = \frac{1}{30} \sum_{i=1}^{30} X_i$.

(i) What is the approximate distribution of $\bar{X}$ (give the mean and variance - these will not necessarily be numbers), and given a reason for your answer?

(ii) Make sketch of the distribution of the population. Over your sketch make a sketch of the (density) distribution of $\bar{X}$.

(iii) Suppose that the population mean is $\mu = 5$. Find the probability that the sample mean $\bar{X}$ is greater than 6.5.