STAT 2507   Solution-Assignment # 4 (Chapters 7 and 8)   Fall 2013

Total of marks=100. Marks for each question are given in [ ]

Part I. Lab questions. Use only the blanks left to answer lab questions. DO NOT submit any plot or data that you are asked to generate.

1. **Central limit theorem (CLT) at work** (You can open a new Minitab worksheet, simply by typing *new*). Generate and store in columns c3-c1002 80 horizontal samples, each of size $n = 1000$, from exponential distribution with mean $\mu = 12$ as follows:
   random 80 c3-c1002;
   expo 12.
   Note This may take a few moments as you are generating 1000×80=80,000 values!
   Create and store in column c1 the 80 values of $\bar{x}$ based on the 80 horizontal samples, each of the same size $n = 1000$ as follows:
   rmean c3-c1002 c1
   [**2**] **a.** Generate the boxplot of the first sample c3. According to the median position and/or the outliers, what can you conclude about the shape of this data set?**skewed to the right**
   [**4**] **b.** Use *desc* command to find sample mean**11.32** and median**6.58** of c3. Do they confirm your diagnostic for the shape above?**yes since mean is larger than median**
   [**2**] **c.** Generate the histogram for the data in column c1. What can you conclude about the shape of data in c1?**fairly symmetric**
   [**3**] **d.** Use *desc* to find sample mean**11.960** and sample standard deviation**0.358** of c1. Are they close to 12 and $12/\sqrt{1000}$?**yes** Why?**since sample mean will have mean $\mu = 12$ and std $12/\sqrt{1000}$= 0.379**
   [**3**] **e.** Do the values of the mean and median of c1 confirm your conclusion in (**c.**) about the shape of data in c1?**yes** Why?**The median is 11.918: very close to the mean=11.960**

2. **Confidence interval (CI) for a population mean:** We want to build 120 confidence intervals (CIs) with confidence level $(1 - \alpha)100\% = 95\%$ for the mean μ of a Poisson distribution via the following steps:
   Step 1. Open a new worksheet. Generate and store in columns c6-c505 120 samples of size 500 each, from Poisson with parameter $\mu = 5$ as follows:
   random 120 c6-c505;
   poisson 5.
   Step 2. Use columns c4 and c5 to store respectively the means and the standard deviations of the 120 horizontal samples you generated in step 1, as follows:
   rmean c6-c505 c4
   rstd c6-c505 c5
   Step 3. Store the lower bound and the upper bound of each of your 95% CIs in c2 and c3 respectively by typing successively:
   *let c2=c4-1.96*c5/sqrt(500)*
   *let c3=c4+1.96*c5/sqrt(500)*
   Step 4. Then create a column c1 containing 1 or 0 according to whether the corresponding interval [c2 , c3] covers μ or not, by typing in the logical function:
   let c1=(c2 <= 5 and c3 >= 5)

Finally sum up the entries of column c1 to find out how many CIs did cover the value $\mu = 5$ by typing:

tally c1

[3] **a.** How many confidence intervals that did contain the true value $\mu = 5$?**111**

[3] **b.** How do you compare this number to the confidence level 95%?**(111/120)100%=92.5%. So the two percentages are very close to each other.**

Part II. Long-answer questions

1. Suppose a random sample of $n = 36$ observations is selected from a population that has normal distribution with mean 106 and standard deviation 12.

   [5] **a.** Give the mean and the standard deviation of the sample mean $\bar{X}$.

   [5] **b.** Find the probability that $\bar{X}$ exceeds 108.

   [5] **c.** Find the probability that the sample mean deviates from the population mean by less than 3.

   Solution:

   (a) $E(\bar{X}) = 106$ and S.D. of $\bar{X} = \frac{12}{\sqrt{36}} = 2$

   (b) Since $n = 36 > 30$, we can use the CLT (central limit theorem) and approximate the sampling distribution of $\overline{X}$ with normal distribution and get:

   $$P(\overline{X} > 108) \approx P(z > \frac{108 - 106}{2}) = P(z > 1) = 1 - P(z \leq 1) = 1 - 0.8413 = 0.1587$$

   (c) Here we are looking for $P(|\overline{X} - \mu| < 3)$. Although we know that $\mu = 106$, but we do not require it to solve problems like this one. See below to figure out what I mean! Applying CLT we get:

   $$\begin{aligned} P(|\overline{X} - \mu| < 3) &\approx P(|z| < \frac{3}{2}) \\ &= P(-1.5 < z < 1.5) \\ &= P(z < 1.5) - P(z < -1.5) \\ &= 0.8664 \end{aligned}$$

2. By statistics, faculty with rank of assistant professor (AP) finishing their 2nd year of employment at a higher education institution in Ontario earn an average of \$ 68,500 per year with a standard deviation of \$3500. In an attempt to verify this salary level, a random sample of 49 AP with 2 years of experiment was selected from a personnel database for all higher education institutions in Ontario.

   [5] **a.** Describe the sampling distribution of the sample mean $\bar{X}$ of the average salary of these 49 AP.

   [5] **b.** Within what limit would you expect the sample mean to fall with probability .95

   [5] **c.** Obtain the probability that $\bar{X}$ is greater than 70,000.

   Solution:

   (a) As $n = 49 > 30$, then by the CLT, $\overline{X}$ can be considered as having (approximately) normal distribution with mean = 68,500 and standard deviation $\sigma/\sqrt{n} = 3500/\sqrt{49} = 500$ dollars.

   (b) Since $n = 49 > 30$, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ has approximately standard normal distribution, with probability .95, we would expect $\overline{X}$ to fall in

   $$\mu \pm 1.96\sigma/\sqrt{n} = 68500 \pm 1.96 \times 500 = [67520, 69480]$$

(c) $P(\overline{X} > 70,000) \approx P(z > \frac{70,000-68500}{500}) = P(z > 3) = 1 - P(z \leq 3) = 1 - 0.9987 = 0.0013$

3. **[10]** In a report on why e-shoppers abandon their online sale transactions, a study found that "pages took me too much time to load" and "site was too confusing to me so that I couldn't find the product" were the two main complaints heard most often. Based on 40 customers' responses, the average time to complete online order was 4.3 minutes and the standard deviation was 2.6 minutes. Construct an 80% confidence interval for μ the average completion time for an online order.
Solution: The sample mean and sample standard deviation are respectively $\overline{X} = 4.3$ and $s = 2.6$. The sample size is $n = 36 > 30$. We want an 80% confidence interval (CI). Hence $\alpha = 0.2$ and $z_{\alpha/2} = z_{0.1} = 1.28$ (from normal table). So, the desired CI is

$$\overline{X} \pm 1.28\frac{2.6}{\sqrt{36}} = 4.3 \pm 1.28 \times 0.43 = [3.75, 4.85]$$

4. **[10]** In a poll of 1000 randomly selected adults, 250 indicated that movies are getting better. Construct a 95% confidence interval for the overall proportion of adults who say that movies are getting better.
Solution: We have $n = 1000$; $\hat{p} = \frac{250}{1000} = 0.25$; $1 - \hat{p} = .75$, and $\alpha = 0.05$. Therefore $z_{\alpha/2} = 1.96$. The 95% CI for p is then given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{250}} = 0.25 \pm 1.96\sqrt{\frac{.25(1-.25)}{250}} = [0.1963, 0.3037]$$

5. **[10]** (choosing sample size) Suppose you wish to estimate the mean pH of rainfalls in an area that suffers heavy pollution due to the discharge of smoke from a power plant. Previous studies showed that the standard deviation is in the neighborhood of .5 pH, and you wish your estimate to lie within .1 of the unknown mean μ, with probability .90. Approximately how many rainfalls must be included in your sample?
Solution:
The sample size n must satisfy (with B being the bound on the margin of error)

$$n \geq (\frac{\sigma z_{\alpha/2}}{B})^2 = (\frac{(.5)(1.64)}{.1})^2 = 67.24$$

and therefore n can be taken equal 68.

6. An experiment was conducted to test the effect of a new drug on a viral infection. The infection was induced in 100 mice, and the mice were randomly split into two groups of 50. The first group, the *control group*, received no treatment for the infection. The second group received the drug. After a 30-day period, the proportions of survivors, $\hat{p}_1$ and $\hat{p}_2$, in the two groups were found to be 0.36 and 0.60, respectively.
[6] Use a 95% confidence interval to estimate the actual difference in the cure rates, i.e. $p_1 - p_2$, for the treatment versus the control groups.

[4] Based on this confidence interval can you conclude that the drug is effective? Why?
Solution: It is important to note that here the 95% confidence interval for $p_1 - P_2$ is in fact the acceptance region of the test $H_0 : p_1 - p_2 = 0$ versus $p_1 - p_2 \neq 0$ at level $\alpha = 0.05$. And the confidence interval is

$$\Big(\hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\Big) = \Big(0.36 - 0.6 \pm \sqrt{\frac{(0.36)(0.64)}{50} + \frac{(0.6)(0.4)}{50}}\Big) = \Big(-0.4301, -0.0498\Big).$$

Since 0 is not in the confidence interval (i.e., acceptance region) we reject H_0.

7. **[10]** Assume that you have a 95% C.I. for a population mean based on a sample of size $n = 30$. If you would wish to have a C.I. of the same confidence level but with a length which is one half of the one you already have, then what the sample size n should be?
   Solution:
   The length of the z-interval for the population mean μ is $\ell = 2\frac{\sigma}{\sqrt{n}} z_{\alpha/2}$. If we replace n by $4n$, then length becomes $\frac{\ell}{2}$. Hence, in order to reduce the length of an interval based on a sample size n by half, we need to have a sample of size $4n$. As a result, in this problem we need to have a sample of size $4(30) = 120$.