PUBH 511 Inferential Biostatistics

&

MSMR 510 Biostatistics and Epidemiology

Assignment #2

Due Date: **12:59 pm** on **Wednesday, November 17th, 2021**

Problem set #1

Your job is to determine the sampling distribution of a mean of N waiting times across all hospitals. You are aware that the population distribution is right skewed where waiting for 32 hours is very rare.

In this experiment, you will investigate the probability distribution of a sample mean for samples of size =2, 5, and 25 number of waiting times. In doing so, you will observe the Central Limit Theorem (see Lecture Notes) in action and learn to explain it to a lay person. The distribution of a sample statistic (here the sample mean) is referred to as a "sampling distribution."

i. **Getting started: Go to the following web site:**

http://onlinestatbook.com/stat_sim/sampling_dist/index.html

You will use the applet to generate random numbers from a true population distribution, here the uniform distribution on the interval (0, 32)

When the page comes up, wait for the "Begin" button to appear. Click "Begin" to open the site. When the site ("applet") opens, notice that there are 4 sections with graphs labeled "Parent Population", "Sample Data", "Distribution of Means" and a fourth graph that is not labeled. Click on the menu at the upper right ("Normal") in the first section, and select "Skewed." Notice how the shape of the graph labeled "Parent Population" changes. Below, where it says "N=5" (under "Mean"), click the menu and change the selection to "N=2." Now, you are ready to do the next step ii.

PLEASE NOTE: In this applet, N refers to sample size which we denote in class as n.

ii. **Means of samples of size n=2:**

Click "Animated Sample." This will generate a sample of two random waiting times that will appear in the second graph labeled "Sample Data." Notice that the mean of the two times is given to the left of that graph, and a small "mass of probability" is added at its value in the third graph labeled "Distribution of Means," just below.

Repeat the step above for a total of 20 times by continuing to click "Animated Sample" 19 more times to form a histogram of mean values.

Describe the shape of the histogram of means recorded by the applet in the third graph. Record and report the mean and standard deviation of the distribution of means for n=2.

iii. **Means of samples of size n=5:**

Now, go to the top section and click “Clear lower 3". Then, click the menu “N=2" in the third section and change it to “N=5". Repeat the procedure using “Animated Sample” to obtain 20 means of 5 waiting times.

Again, describe the shape of the distribution of means given by the applet in the third graph. Record and report the mean and standard deviation of the distribution of means.

iv. Means of samples of size n=25:

Go to the top section and click “Clear lower 3" one more time. Then change “N=5" to “N=25" in the third section. Repeat the procedure using “Animated Sample” to obtain 20 means of 25 waiting times.

Once more, describe the shape of the distribution of means given by the applet. Record and report the mean and standard deviation of the distribution of means.

v. The population distribution from which you have been sampling has mean=8.08 waiting hours and variance 38.69. Complete the table below with the estimated means and standard deviations of the sampling distribution that you obtained from your 20 replicates in parts ii-iv above and with the known theoretical values given by the Central Limit Theorem.

Sample Size (n)	**Observed Statistics for 20 Replicates**		**Theoretical Values for Infinite Replicates**	
	Mean	**Standard Deviation**	**Mean**	**Standard Deviation**
2				
5				
25				
100	**NA**	**NA**		

vi. In a few sentences, explain how far sample means tend to be from the true population mean and how the deviation depends on the population variance and sample size. State the Central Limit Theorem in your own words.

Problem set #2

A large clinical trial carried out by the Radiation Therapy Oncology Group in the United States looking at the treatment of carcinoma of the oropharynx (Source: The Statistical Analysis of Failure Time Data, by JD Kalbfleisch & RL Prentice, (1980), Published by John Wiley & Sons). Patients entering the study were randomly assigned to one of two treatment groups, radiation therapy alone (standard) or radiation therapy together with a chemotherapeutic agent (test). One objective of the study was to compare the two treatment policies with respect to patient survival.

The collected data is saved in the file “pharynx.xls” or “pharynx.RDATA” (this data set is only an extract of the original data).

List of variables:

id	patient identification number
sex	patient sex
tx	treatment: standard or test
age	patient age in years at time of diagnosis
time	survival time in days from day of diagnosis

A. Suppose you were told that parameters for survival time for a population of patients diagnosed with carcinoma of the oropharynx regardless of a treatment type are

μ =the population mean survival time = 660 days

σ =the population standard deviation survival time= 480 days

i. Describe the theoretical sampling distribution of mean survival time of sample size n=80 (shape, central tendency, variability)

ii. How would the sampling distribution change if the sample size was not 80, but 20?

iii. Now, summarize survival time variable of a sample of 80 observations in the data to find a sample mean and sample standard deviation

Loading the dataset:

File->Import dataset -> *From Excel* ->*Browse* ->*Import*

Or you can type the commands below:

```
library(readxl)
pharynx <-read_excel("C:/Location/WHERE/YOU/SAVED/pharynx.xls")
View(pharynx)
```

Do not forget to attach the dataset:

```
attach(pharynx)
```

Now summarize the time variable:

```
summary(time)
sd(time)
```

What is the probability of observing the sample mean 400 or less than that of a sample size n= 80?

iv. What is the probability of observing the sample mean of a sample size n= 80 between 500 days and 700 days?

B. Now, suppose you do NOT know the true population parameters only sample statistics from section iii.

v. Please calculate 90%, 95% and 99% confidence intervals for the true population survival time mean based on the sample of size n=80.

vi. Similarly, using the sample statistics from section iii, please calculate 90%, 95% and 99% confidence intervals for the true population survival time mean based on the sample of sizes n=121 (Z-distribution can be used), n=30 and n=15. Write your calculations in Table 1 below:

Samples with different sizes	90% CI for μ	95% CI for μ	99% CI for μ
n=121 (Z-table)			
n=80			
n=30			
n=15			

vii. Based on the results in sections v and vi, comment on differences that you observe among CI's by sample sizes. Please provide explanations for the dissimilarities you observed.

viii. Do you believe the sample distribution of survival time from section iii follows normality? Write your justifications. You can use following commands to come to your conclusion:

```
hist(time)
qqnorm(time)
```

ix. Assume that the sample survival time distribution can be reasonably approximated by a normal distribution. Calculate intervals of survival time values for middle 50%, 80%, 90% and 98% of observations.

x. Contrast your calculations assuming normality from section ix with the actual percentile survival time values by filling out the Table 2 below:

Obtaining actual survival time values by percentiles:

```
quantile(time, probs = c(0.01, 0.05, 0.1, 0.25, 0.50, 0.75,
0.90, 0.95, 0.99))
```

Table 2: Comparing actual data vs assuming Normality

Percentiles (in the brakes, under normality assumption)	Actual data	Assuming Normality
1% (lower limit of the interval containing middle 98%)		
5% (lower limit of the interval containing middle 90%)		
10% (lower limit of the interval containing middle 80%)		
25% (lower limit of the interval containing middle 50%)		
50% (mean)		
75% (upper limit of the interval containing middle 50%)		
90% (upper limit of the interval containing middle 80%)		
95% (upper limit of the interval containing middle 90%)		
99% (upper limit of the interval containing middle 98%)		

xi. Fill out the table in Excel file “Graph” using results in the section x. You should notice two lines on the graph: one for actual data and another assuming normality. You would see that under normality assumption survival time values tend to follow straight line while actual data points may not follow (if it is not normal). If actual data for survival time is normally distributed, then two lines must be overlap with each other whereas dots aligned on the normal curve line. A similar Quantile-Quantile (Q-Q) plot can be obtained using R-software command: `qqnorm(time)` . Please comment whether the sample distribution approximates a normal distribution.

C. Calculating confidence interval & hypothesis testing

xii. Calculate one-sided 95% confidence interval (upper bound) for the true population survival time among male patients.

```
summary(time[sex=="male"])
sd(time[sex=="male"])
```

xiii. A previous study found that average survival time among male patients was 650 days. Test the hypothesis whether the true population survival mean for male patients is statistically significantly less 650 days, at significance level 0.05.

xiv. Comment whether or not you can draw conclusion about whether or not the true population survival mean for male patients is less 650 days only based on confidence interval calculated from the section xii.

xv. Calculate two-sided 95% confidence intervals for the true population survival time by sex. Verify your calculations with R output:

```
t.test(time[sex=="male"])
t.test(time[sex=="female"])
```

xvi. Name assumptions for constructing confidence intervals in the section xv. Check and provide evidence whether or not these assumptions are met.

xvii. Based on the results obtained from section xv, comment whether or not you believe the true population survival time means are different from each other.