ECO240 R- Homework 1 <u>[Due Date: April 7, Thursday at 16:00]</u>

<u>Submit your report to my office.</u>

<u>No late HW will be accepted. Even partially copied HW will get 0 point.</u>

Task1: Random Sampling
Task2: Find critical values of t-distribution
Task3: Find critical values of chi-squared distribution
Task4: Calculate a Confidence Interval Estimate of Population Mean when σ is known.
Task5: Calculate a Confidence Interval Estimate of Population Mean when σ is unknown.
Task6: Calculate a Confidence Interval Estimate of Population Variance
Task7: Calculate a Confidence Interval Estimate of Population Proportion
Task8: Calculate a Confidence Interval Estimate of Difference in Population Means

*Tasks 1~9 will be done by writing scripts in RGui (the main window of R) script window, or in RStudio.
*Some tasks can also be done by using "*Rcmdr*" package. First download the package using one of CRAN sites. On RGui window, find Packages tab -> load package -> select "rcmdr". R Commander window pops up.
*Download **housingdata.csv** for Task 1. **finalscores.csv** for Task 4,5,6,7,8 and 9.

*Reporting: Submit scripts written in Editor window, outputs in Console window and observations/discussions of your own.

<u>Task 1. Random Sampling</u>

We are going to randomly sample n = 100 observations from N = 10665 of housingdata.csv.

a. In order to sample n out of N observations, the simplest way is to select 100 numbers randomly out of 10665 integers. The command to do so is by typing the following to the RGui, R Console window.

   ```
   >sample(x, size, replace = FALSE, prob = NULL)
   ```

   where, x is the total number of observations, size is the sample size, replace means if you want to sample with or without replacement (If you do not want replacement, set it as FALSE, if you want replacement, set it as TRUE), prob is the type of probability weight for more complex sampling.

   Write the code with x of your data, and see if it returns 100 randomly sampled integers.

b. Create a data file with 100 observation IDs you just sampled. Import the data file and create a histogram based on ID values by setting the number of bins as 11. Comment on any pattern(s) you observe. Report the 100 IDs selected and the created histogram.

<u>Task2: Find critical values of t-distribution</u>

By using "`qt`", you can find the critical values from t-distirubtion The basic syntax is

```
qt(p, df, lower.tail = FALSE)
```

where p is the probability, df is the degrees of freedom, lower.tail=FALSE means that p is upper tail probability. For example, if we want to obtain the t-value for $P(t > t_{10,0.05})$ = 0.05 [= try to obtain the t-value that upper tail probability is 0.05 given df = 10], we write a script as

```
qt(0.05, 10, lower.tail = FALSE)
```

as the output, we obtain

```
> qt(0.05, 10, lower.tail = FALSE)
[1] 1.812461
```

Due to the symmetry of t-distribution, if we change upper probability to lower probability by setting lower.tail = TRUE, we get $-t_{10,0.05}$ = -1.812461 as below.

```
> qt(0.05, 10, lower.tail = TRUE)
[1] -1.812461
```

Task2-1 :a. Find the t value when 5% of all values are above this value, given df = 20.
b. Find the t value when 5% of all values are below this value, given df = 20.
c. Find the t values when 95% of all values are between these values, given df = 20.

Task 2-2: a. Find the t value when 10% of all values are above this value, given df = [your choice].
b. Find the t value when 80% of all values are below this value, given df = [your choice].

Task 2-3: [Use Rcmdr package to answer this question. Distribution -> Continuous Distribution ->t distribution]
Observe the change in the density function as you change df. Plot figures for df=5, df=10,df=100 and comment.

Task3: Find critical values of chi-squared distribution

By changing “qt” to “qchisq”, you can find the critical value from Chi-squared distribution. The basic syntax is

```
qchisq(p, df, lower.tail = FALSE)
```

Task3-1 :a. Find the $\chi^2_{v,\alpha}$ value when 5% of all values are above this value, given df = 20.
b. Find the $\chi^2_{v,\alpha}$ value when 5% of all values are below this value, given df = 20.
c. Find the $\chi^2_{v,\alpha}$ values when 95% of all values are between these values, given df = 20.
Task 3-2: Select a degrees of freedom of your own and repeat Task 3-1.

Task 3-3: Using Rcmdr, plot some figures changing df values. Comment on the changes in the shapes of density functions.

Task4: Calculate a Confidence Interval Estimate of Population Mean when σ is known.

We will derive confidence interval of population mean when population standard deviation is known under this task. We use FinalScore variable from finalscores.csv file uploaded on our course web. **This is a population data** of size 73 (N = 73). Then import it by typing the following in RGui Editor window. (If you have any problem, suspect that you may be having "," vs. "." conflict in your file. To avoid the problem, try saving the csv file without opening before importing to R program.)

```
finalscores <- read.table("d:/finascores.csv", header=TRUE,sep=",")
```

*Make sure to change the location of the directory where the file is saved.

Inside the parentheses, change the location of the file to your local directly. Make sure that you use “/” instead of “\” as the separator. You can observe the data by simply typing

```
finalscores
```

It contains 3 variables: ID, final2013, final2014

We will use final2014 variable in this question. Let's define it as:

```
final2014= finalscores[,3]
```

# Computing margin of error (m.e.) by finding z value satisfying P(Z<Z)=0.975. Since it's 95% C.L., #1-α = 0.95, α=0.05, α/2=0.025.* `qnorm(0.975)` *gives the z value. Multipy it by* $\sigma/\sqrt{n}$.

```
m.e.<-qnorm(0.975)*sigma/sqrt(n)
```

# lower confidence limit

```
LCL<-x_bar-m.e.
```

# upper confidence limit

```
UCL<-x_bar+m.e.
```

Our confidence interval is computed as [LCL UCL] or [4.1234   5.8765]

Task4-1: Import data file.
Task4-2: Compute population mean (mu) ,population standard deviation (sigma) and population proportion p (x<40).
Task4-3: Randomly sample 20 observations from the population data, generate and import .csv file of the generated sample data.

```
# take a random sample of size 50 from a dataset finalscores
sample_final <- finalscores[sample(1:nrow(finalscores), 20,replace=FALSE),]
```

Task4-4: Compute sample mean (x_bar) and standard deviation (s) of the sample.
Task4-5: Compute 95% and 99% confidence intervals for population mean.
Task4-6: Comment on the ranges of C.I.s for different confidence levels.
Task4-7: Comment on the relationship between the found C.I.s and actual population mean from Task4-2.
*For Tasks 4-2,4-3,4-4, and 4-5, copy-paste the scripts and outputs to your report.

Task5: Calculate a Confidence Interval Estimate of Population Mean when σ is unknown.

For this task, we use the same sampled data created in Task4-3. Do the necessary adjustments and complete the following tasks.

Task5-1: Compute 95% and 99% confidence intervals for population mean.
Paste your scripts and output to the report.
Task5-2: Comment on the relationship between the found C.I.s and actual population mean from Task4-2.
Task5-3: Comment on the similarities/differences of the results from Task4-5 and Task5-1.

Task6: Calculate a Confidence Interval Estimate of Population Variance

For this task, we use the same sampled data created in Task4-3.

Task6-1: Compute 95% and 99% confidence intervals for population variance.
Paste your scripts and output to the report.
Task6-2: Comment on the relationship between the found C.I.s and actual population variance from Task4-2.

Task7: Calculate a Confidence Interval Estimate of Population Proportion

For this task, we use the same sampled data created in Task4-3.

Task7-1: Sort the data and compute the proportion of FinalScore<40 (Proportion of the students scored less than 40.)
Task7-2: By setting the computed proportion of FinalScore<40 as p̂, 95% and 99% confidence interval estimates for population proportion. Although np(1-p) may not exceed 5, assume normality.
Task7-3: Compare the population proportion calculated in Task 4-3 with the ranges derived for Task 7-2. Comment.

Task 8: Calculate a Confidence Interval Estimate for the Differences between Population Means

The objective of this task is to calculate the confidence interval for the difference between the population mean scores of ECO240 conducted in 2013 (2nd column) and 2014 (3rd column).

Task8-1: Compute population means and population standard deviations for each variable.

Task8-2: Randomly sample 20 observations from the population data, generate and import .csv file of the generated sample data.

Task8-3: Compute sample means (x_bar) and standard deviations for each variable.

Task8-4: Compute 95% and 99% confidence intervals for population mean.

a) Independent sample, population variances known case.

b) Independent sample, population variances unknown, but assumed to be equal.

c) Independent sample, population variances unknown.

Task 8-5: Compare the results from Task 8-4 a, b, c and the size of $\mu_x - \mu_Y$ and comment.