ECO240 R- Homework 1 [Due Date: April 7, Friday at 16:00]

- **Submit a hard copy to my office before the due. No late homework will be accepted.**
- **Submit "*Contribution Paper*" and "*Honor Code*" forms *signed* by all the group members.**
- **Your group can be up to 4 students.**
- **Any copied/being copied HW will get ZERO point. No negotiation.**
- **If suspected, I reserve the right to invite you for the further investigation.**
- **Please read "Honor Code" document carefully to understand what COPY means.**

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

*Use RStudio for all the tasks. You may use "*Rcmdr*" for Task 2 and 3.
*To use "*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.

Task 0. Selection of Data

Task 0-1: Find data set to be used throughout this homework. The data should contain at least 100 observations with at least one continuous variable. You may find your data set from TUIK, OECD database, or R packages such as AER, Ecdat, EconDemand, erer… [https://cran.r-project.org/web/packages/available_packages_by_name.html]

Task 0-2: Select one variable (continuous variable) from the data set. Report: Variable description and the population (who/what are the population) of data.

Task 0-3: Check if the selected variable is normally distributed or not. If your variable has very skewed distribution, select different variable with relatively symmetric distribution.

Task 0-4: Create at data file with two variables, ID and the selected variable (to be used in Task 1).

Task 1. Random Sampling

Task 1-1. Randomly sample 30 observations from your data set form Task 0-4.

a. In order to sample n out of N observations, the simplest way is to select 30 numbers randomly out of N integers. The command to do so is by typing the following code.

   ```
   >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.

Task 1-2: Create a data file with 30 observation IDs you just sampled. Create a histogram based on ID values to see the distribution of the selected IDs of the observations. Comment on any pattern(s) you observe. **Report the 30 IDs selected and the created histogram**.

* You will use this random sample (n=30) for the remaining of this homework (Tasks 2~7).

Task2: Find critical values of t-distribution

By using "qt", you can find the critical values from t-distribution 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 = 29.
b. Find the t value when 5% of all values are below this value, given df = 29.
c. Find the t values when 95% of all values are between these values, given df = 29.

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 = 29.
b. Find the $\chi^2_{v,\alpha}$ value when 5% of all values are below this value, given df = 29.
c. Find the $\chi^2_{v,\alpha}$ values when 95% of all values are between these values, given df = 29.

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. Use the same sampled data from Task 1 for this task.

# Computing margin of error (m.e.) for 95% C.I. 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 σ/√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].

Task4-1: Import data file.
Task4-2: Compute population mean (mu) and population standard deviation (sigma) by using your original data (before random sampling) used in Task 1.
Task4-3: Compute sample mean (x_bar) and standard deviation (s) of the sample data from Task 1 .
Task4-4: Compute 95% and 99% confidence intervals for population mean.
Task4-5: Comment on the relationship between the found C.I.s and actual population mean from Task4-2. Are the C.I.s containing the population mean?
*For Tasks 4-2,4-3,4-4, and 4-5, paste the scripts and outputs to your report.

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

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 Task4-4 and Task5-1.

Task6: Calculate a Confidence Interval Estimate of Population Variance

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

Task7-1:Calculate the mean of the sample data.
Task7-2: Compute the proportion of data X < mean.
Task7-3: By setting the computed proportion of X<mean as $\hat{p}$, find 95% and 99% confidence interval estimates for population proportion. Although np(1-p) may not exceed 5, assume normality.
Task7-4: Calculate population proportion using the ORIGINAL data (before sampling) using the same criteria (X< sample mean). Compare the calculated population proportion with the ranges derived for Task 7-3. Comment.