**STAT371** **Discussion #5**

**I**. 1-$\alpha$ Confidence Interval (CI) for Sample Mean $\overline{X}$

1. If population SD $\sigma$ is known, then $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = Z \sim N(0,1)$

(1) The 1-$\alpha$ CI is $(\overline{X} - Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}})$, where $Z_{\alpha/2}$ is the $(1-\frac{\alpha}{2})$ quantile of standard normal distribution

(2) There are two ways to find $Z_{\alpha/2}$

- Use Z table
- Use **qnorm (**$1-\frac{\alpha}{2}$**, 0, 1)**

Example1: when $\alpha$=0.05, the CI calculated is a 95% CI. Then $Z_{\alpha/2}$ is 1.96, which is the 97.5% quantile of standard normal distribution

2. If population SD $\sigma$ is not known, then $\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$

(1) The 1-$\alpha$ CI is $(\overline{X} - t_{\alpha/2,\text{n-1}} \times \frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2,\text{n-1}} \times \frac{s}{\sqrt{n}})$, where $t_{\alpha/2,\text{n-1}}$ is the $(1-\frac{\alpha}{2})$ quantile of the t distribution with n − 1 “degrees of freedom”.

(2) There are two ways to find $t_{\alpha/2,\text{n-1}}$

- Use t table
- Use **qt(**$1-\frac{\alpha}{2}$**, n-1)**

(3) **t.test()** in R can also calculate the confidence interval , see IV below

**II**. Structure of Excel Data File

1. Columns and rows
   (1) Columns: each column indicates a variable
   (2) Rows: each row is a complete record of an observation

2. Keep track of as many things as are feasible when you record the data
   - subject ID or run ID
   - date
   - treatment
   - various responses/outcomes
   - in animal experiments, always keep track of the sex of the animals
   - Put something like "NA" in cells to indicate missing values (rather than leave them blank); this helps to make errors more obvious

   I often see data of the form "bad.csv", which is generally awkward to work with. Much better is "good.csv".

**III**. How to get excel data file into R

Step1: Save an Excel file as a comma-delimited file with extension .csv
Step2: Load the CSV file into R using **read.csv()**

**good<-read.csv('good.csv', header=T)**

**IV**. How to calculate confidence intervals in R, using t.test()

(1) t.test() is a function used to perform t-test. It generates several useful results including confidence interval that we want.

(2) t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0,

```
paired=FALSE, var.equal = FALSE, conf.level = 0.95, ...)
```

(3) Many options, but at this time, we only use **t.test(x, conf.level=)**
   - x: name of the variable for which you want to find confidence interval
   - conf.level= : confidence level of the interval, 0.95 is 95% and so on. Default is 95%.

Example1 95% CI: **t.test(good$response)**
Example1 90% CI: **t.test(good$response, conf.level=0.90)**
$ here means the response variable is from good data file.

**V**. Subsets of the data

1. How to pull out a particular column or columns
   Example1, pull out column #2 from good data file: **good1<-good[,2]**
   Example2, pull out column #2 - #4 from good data file: **good2<-good[,2:4]**

2. How to pull out a particular row or rows

Example3, pull out row #8 from good data file: **good3<-good[8,]**
Example4, pull out row #8 - #15 from good data file: **good4<-good[8:15,]**

3. How to pull out subsets of the data
   Example5, try **good5<-good[8:15,2:4],** what can you get?
   Example6, take a subset with treatment=A
   **treatA<-good[good$treatment =='A',]**

4. Pull out subsets using columns combining and rows combining functions
   Example7, take a subset with column id and column response
   **good7<-cbind(good$id, good$response)**
   Example8, take a subset with rows 1-5 and rows 11-15
   **good8<-rbind(good[1:5,], good[11:15,])**