STAT 3155—-Introduction to R

- Addition and subtraction

```
> 4+3
[1] 7
> 5-2
[1] 3
```

- Multiplication and division

```
> 10*1.5
[1] 15
> 30/6
[1] 5
```

- Exponents

```
> 4^3
[1] 64
> 4^-3
[1] 0.015625
```

- storing and manipulating

```
> x=5
> y=7
> x+y
[1] 12
> x*y
[1] 35
```

- Combine

```
> c(-1, 2, 5)
[1] -1  2  5
> c("a","a","b", "c")
[1] "a" "a" "b" "c"
```

- Sum and Mean

```
> sum(c(1,3,5))
[1] 9
> mean(c(-1,4, 10))
[1] 4.333333
```

- Variance and standard deviation

```
> var(c(1,5,-2,7))
[1] 16.25
> sd(c(1,5,-2,7))
[1] 4.031129
```

- Minimum and Maximum

```
> min(c(-1,-5,2,10))
[1] -5
> max(c(-1,-5,2,10))
[1] 10
```

- Covariance and Correlation

```
> x<-c(-1,2,3,9)
> y<-c(2,0 ,4,7)
> cov(x,y)
[1] 10.25
> cor(x,y)
[1] 0.8186005
```

- Combine as columns

```
 > x<-c(-1,2,3,9)
> y<-c(2,0 ,4,7)
> cbind(x,y)
      x y
[1,] -1 2
[2,]  2 0
[3,]  3 4
[4,]  9 7
```

- Combine as rows

```
> rbind(x,y)
  [,1] [,2] [,3] [,4]
x   -1    2    3    9
y    2    0    4    7
```

- summary function

```
> x<-c(1,5,-8, -4, 2, 6)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-8.0000 -2.7500  1.5000  0.3333  4.2500  6.0000
```

- Probability distributions
  Example 1: When working with R we often want to make probability statements based on a distribution. For example suppose X has a normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 4$.

  a) Find $P(X < 14)$

  b) Find the value of x if $P(X < x) = 0.9772$.

  c) Generate a random sample of size $n = 10$ from this distribution

  Answers: a)

```
> pnorm(14,10,4)
[1] 0.8413447
```

  b)

```
> qnorm(0.9772,10,4)
[1] 17.99631
```

  c)

```
> rnorm(10,10,4)
 [1]  8.233378 16.357944  8.616639 15.035167 11.757097 10.498963  7.146407
 [8]  3.446712 17.807686  2.941490
```

- The general name structure is
  - pname calculates the cumulative distribution function at the input value
  - qname calculate the quantile at the input probability
  - rname generates random sample on input size

  Here name is the name of the distribution of interest

One sample t-test

- Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean 0 and variance σ^2. where σ^2 is unknown.
- Goal: Test $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$. (μ_0 is given)
- The test statistics is
$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$
  where $\bar{X}$ is the sample mean and S is the sample standard deviation and we reject H_0 if $|t| > t^{(n-1}_{[\alpha/2]}$ or if $p-$value $< \alpha$ where $t_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the t-distribution with $n - 1$ degrees of freedom
- A $100(1 - \alpha)\%$ confidence interval for μ is given by
$$\bar{X} \pm t^{(n-1)}_{[\alpha/2]} S/\sqrt{n}$$

- Example suppose we wish to test $H_0 : \mu = 0$ against $H_a : \mu \neq 0$ using $\alpha = 0.05$ and the following data

```
> x<-rnorm(20,0,1)
> x
[1] -0.02683366 -1.01949430  0.99563275 -1.56184620  1.72165151  1.44689548
 [7] -2.18422523 -1.75340520  1.21104040 -0.33665061 -1.67055420 -0.44641791
[13] -0.77499946  0.51894632  0.13233206  0.23850408  0.19045199  0.87856402
[19] -0.09381772 -0.60813365
```

- Using R we do the following

```
> t.test(x)

One Sample t-test

data:  x
t = -0.63072, df = 19, p-value = 0.5357
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.6785086  0.3642727
sample estimates:
mean of x
-0.157118
```

- To test $H_0 : \mu = 1$ against $H_a : \mu \neq 1$, use

```
> t.test(x, mu=1)
```

- The answer is

```
One Sample t-test

data:  x
t = -4.645, df = 19, p-value = 0.0001765
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 -0.6785086  0.3642727
sample estimates:
mean of x
-0.157118
```

- To test $H_0 : \mu = 0$ against $H_a : \mu < 0$. do the following

```
One Sample t-test

data:  x
t = -0.63072, df = 19, p-value = 0.2679
```

```
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
      -Inf 0.2736242
sample estimates:
mean of x
-0.157118
```

Two sample t-test

The two-sample t-test (Snedecor and Cochran, 1989) is used to determine if two population means are equal. A common application is to test if a new process or treatment is superior to a current process or treatment. We need to consider three cases:

1. Case 1. In this case we have independent samples $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ from two populations with means μ_1 and μ_2, respectively and these populations are assumed to have equal variances ($\sigma_1^2 = \sigma_2^2$). We wish to test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu$. The test statistic is

$$t = \frac{\bar{X} - \bar{Y} - 0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

and S_1^2 is the sample variance corresponding to the first sample and S_2^2 is the sample variance corresponding to the second sample, We reject H_0 if

$$|t| > t_{[\alpha/2]}^{(n_1+n_2-2)}$$

or if $p-$value $< \alpha$.

```
> x
 1.08785019  0.09545013  0.87524589 -1.09295503 -0.36975383 -0.32075040
 -0.19797132 -0.36600452  0.42431836 -1.02238611 -1.61256622  1.56872597
 -0.99798478  0.59961134 -0.06927037  0.81386972 -0.10080678  1.47196315
  0.68408470 -1.06014442
> y
 [-0.66572781  1.05056852  0.01157729  0.52405616 -0.66421951 -0.19959536
 1.05770527  0.65120152  0.11028454 -0.79903786 -0.43426068 -0.06548589
-0.49502723 -0.75404563  0.50111347  0.51621344  0.22860289  1.19463425
-0.68436718  0.11205275

> t.test(x,y,paired=FALSE, var.equal=TRUE)

Two Sample t-test

data:  x and y
t = -0.15783, df = 38, p-value = 0.8754
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 -0.5431894  0.4646177
sample estimates:
 mean of x  mean of y
0.02052628 0.05981215
```

2. Case 2, same setting as before but we do not assume that variances are equal. In this case the test statistic is

$$t = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

and we reject H_0 is $|t| > t_{\alpha/2}^{(\nu)}$ or $p-$value $< \alpha$. Here

$$\nu = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}$$

```
> x
 [1]  1.08785019  0.09545013  0.87524589 -1.09295503 -0.36975383 -0.32075040
 [7] -0.19797132 -0.36600452  0.42431836 -1.02238611 -1.61256622  1.56872597
[13] -0.99798478  0.59961134 -0.06927037  0.81386972 -0.10080678  1.47196315
[19]  0.68408470 -1.06014442
> y
 [1] -0.66572781  1.05056852  0.01157729  0.52405616 -0.66421951 -0.19959536
 [7]  1.05770527  0.65120152  0.11028454 -0.79903786 -0.43426068 -0.06548589
[13] -0.49502723 -0.75404563  0.50111347  0.51621344  0.22860289  1.19463425
[19] -0.68436718  0.11205275
> t.test(x,y,paired=FALSE, var.equal=FALSE)

Welch Two Sample t-test

data:  x and y
t = -0.15783, df = 34.306, p-value = 0.8755
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5449776  0.4664059
sample estimates:
 mean of x  mean of y
0.02052628 0.05981215
```

3. Paired t-test. A paired t-test is used to compare two population means where you have two samples in which observations in one sample can be paired with observations in the other sample. Examples of where this might occur are:

   - Before-and-after observations on the same subjects (e.g. students? diagnostic test results before and after a particular module or course).
   - A comparison of two different methods of measurement or two different treatments where the measurements/treatments are applied to the same subjects (e.g. blood pressure measurements using a stethoscope and a dynamap).

4. Example: Suppose a sample of n students were given a diagnostic test before studying a particular module and then again after completing the module. We want to find out if, in general, our teaching leads to improvements in students? knowledge/skills (i.e. test scores). We can use the results from our sample of students to draw conclusions about the impact of this module in general. Let X = test score before the module and Y =test score after the module To test the null hypothesis that the true mean difference is zero, the procedure is as follows:

   (a) Calculate the differences $D_i = Y_i - X_i, i = 1, 2, \ldots, n$,

   (b) Compute $\bar{D}$ and S_D^2 (the sample mean and the sample variance of these differences) and

$$t = \frac{\bar{D} - 0}{S_D/\sqrt{n}}.$$

   Reject $H_0 : \mu_1 = \mu_2$ if $|t| > t_{\alpha/2}^{(n-1)}$ or if $p-$value $< \alpha$.

5. To conduct this test using R do

```
t.test(x, y , mu = 0, paired = TRUE)
```

6. using the data above we get

```
> t.test(x,y,paired=TRUE)

Paired t-test

data:  x and y
t = -0.17052, df = 19, p-value = 0.8664
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5214813  0.4429096
sample estimates:
mean of the differences
            -0.03928586
```