Statistics 601 – Assignment 1

Name: Naiqing Cai

email:ncai5@wisc.edu

1. Numerical methods: Coefficient of Variation

Definition

It is defined as the ratio of the standard deviation to the mean. It is a standardized measure of dispersion of a distribution.

Advantages

The standard deviation of data must always be understood in the context of the mean of the data. In contrast, the actual value of the CV is independent of the unit in which the measurement has been taken, so it is a dimensionless number.

Disadvantages

When the mean value is close to zero, the coefficient of variation will approach infinity and is therefore sensitive to small changes in the mean.

Unlike the standard deviation, it cannot be used directly to construct confidence intervals for the mean.

Example (weather in Madison in winter)

Data： Celsius: [-10, 2, -8, 5, 7, -15]

Fahrenheit: [13, 15, 6, 14, 1, 10]

The sample standard deviations are 9.02 and 5.42, respectively.

The CV of the first set is 9.02/-3.17 = -2.85. For the second set (which are the same temperatures) it is 5.42/9.83 = 0.55

2.(a) (b) (c)

The histograms of sample mean and sample variance from the 100 normal distributed samples when n=10, n=40, n=160 are as followed.

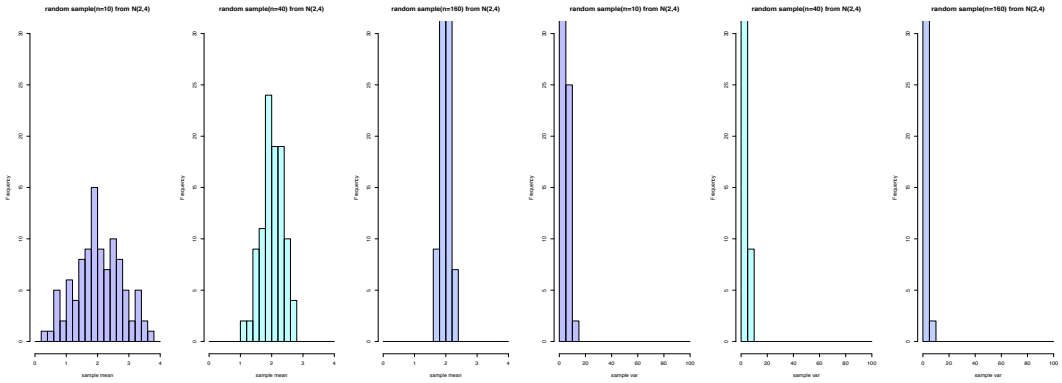


Figure 2.1 sample mean

Figure 2.2 sample variance

2.(d)

From the histograms in (a)–(c), I found that with the increase of n, which is the number of observations each time, the distribution of sample mean and sample variance are more likely to be concentrated, which means the range and the distribution becomes narrower. Specifically, when n=10, the data is close to normal and without outliers. When n=40, mild skewness is acceptable but not outliers. When n=160, it will have strong skewness.

2.(e)

The histograms of sample mean and sample variance from the 100 binomial distributed samples when n=10, n=40, n=160 are as followed.

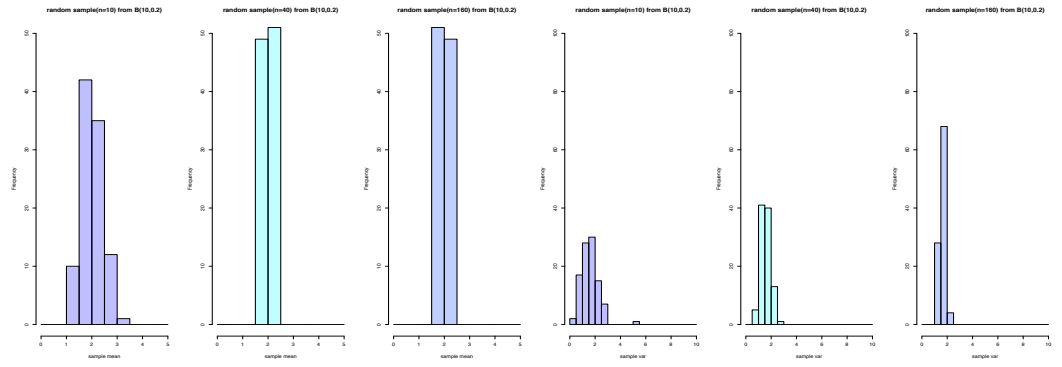


Figure 2.3 sample mean

Figure 2.4 sample variance

From the histograms, it also shows that with the increase of n, the distribution of sample mean and sample variance are more likely to be concentrated, that is the range of both are smaller and the distribution is much narrower.

2.(f)

$E(\bar{Y}) = \mu$

$Var(\bar{Y}) = \frac{1}{n}\sigma^2$

$E(S^2) = \sigma^2$

Compared with my simulations, I find that when n tend to be infinite, the mean of sample mean will be a constant and the variance of sample mean will also be a constant. Also the expectation of sample variance will tend to be the variance the population.

$Var(\bar{Y}) = \frac{1}{n}\sigma^2$ is the variance of sample mean, and $Var(Y) = \sigma^2$ is the variance of samples of the population.

2.(g)

$Y \sim B(m,\pi)$

$E(\bar{Y}) = m\pi$

$Var(\bar{Y}) = \pi(1-\pi)$

$E(S^2) = m\pi(1-\pi)$

$Var(\bar{Y}) = \pi(1-\pi)$ is the variance of sample mean, and $Var(Y) = m\pi(1-\pi)$ is the variance of samples of the population.

3.(a) Hypothesis

$H_0 : \mu = \mu_0$

$H_A : \mu > \mu_0$

Parameters:

μ : where μ is the mean of the population posttest scores minus the mean of the population pretest scores.

μ_0 : equals to 0

3.(b) Graphical check

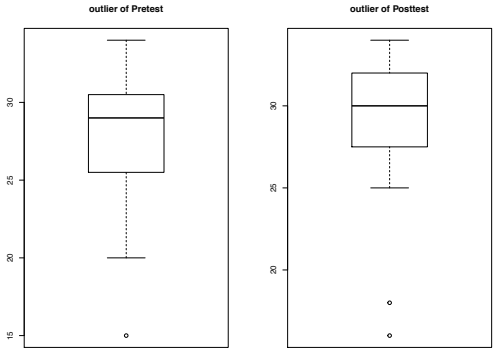


Figure 3.1 outliers

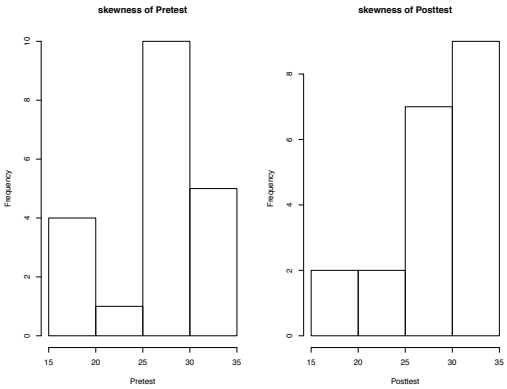


Figure 3.2 skewness

For the outliers, we can conclude from the boxplot that both samples have outliers, but it will be robust to the test.

For the skewness, we can see from the histogram that samples may not from a normal distributed population.

3.(c) T-test (one sample)

Choose one sided t-test with 95% confidence interval

step 1: Hypothesis

$H_0 : \mu = \mu_0$

$H_A : \mu > \mu_0$

step 2: Test Statistics

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim T_{n-1}$$

the standard error is $s/\sqrt{n} = 3.2032/\sqrt{20} = 0.716$

sample mean: $\overline{x} = 1.45$

$$t = \frac{\overline{x} - 0}{s/\sqrt{n}} = 2.0244$$

step 3: Find the p-value

$P(T \geq 2.0244)$ is 0.0286, less than 0.05, so at the 5% level, so we should reject the H_0 . Thus, there is strong evidence that the mean scores of Posttest is greater than that of Pretest. In this way, the training improves listening skills.

3.(d) Confidence Interval

Choose two sided t-test with 90% confidence interval

$$P(-t_{n-1,\frac{\alpha}{2}} \leq T_{n-1} \leq t_{n-1,\frac{\alpha}{2}}) = 1-\alpha$$

$$\mu_X \in [\overline{x} - t_{n-1,\frac{\alpha}{2}} \times (S/\sqrt{n}) \leq T_{n-1} \leq \overline{x} + t_{n-1,\frac{\alpha}{2}} \times (S/\sqrt{n})]$$

$$\mu_X \in [0.212, 2.689]$$

90% confidence interval for the mean increase in listening score due to the intensive training is [0.212, 2.689], which means there are 90% probability that the mean increase in listening score due to the intensive training is in [0.212, 2.689]

3.(e) Find the minimum of n

$$P(T(n) \in [-z_{\alpha/2}, z_{\alpha/2}] \mid D_i \sim N(2,1), i.i.d) \leq \beta$$

$$T(n) = \frac{\overline{X} - \mu_0}{\sigma} = \overline{X} - 0 = \overline{X}$$

$$P(-z_{\alpha/2} \leq \overline{X} \leq z_{\alpha/2}) \leq \beta = 0.2$$

From r code, we can find that $\min n = 2$

4.(a) Boxplot and numerical statistics to summarize the data

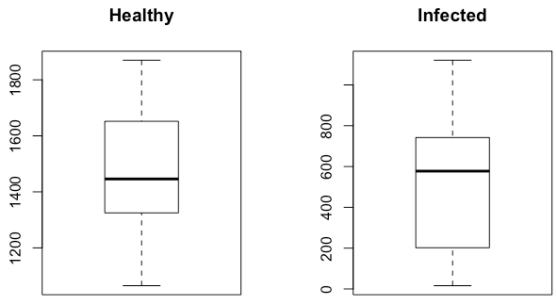


Figure 4.1 boxplot of healthy and infected buds

Conclusions: From the boxplot, we can find that there are no outliers in these two samples.

	Healthy	Infected
median	1446	577.5
mean	1480	549.4286
standard deviation	248.9849	343.4586

Table 4.2 numerical statistics of healthy and infected buds

Conclusions: The virus makes the stem volume of infected 2-year-old seedlings much less than the healthy one and makes the standard deviation of infected one much larger than the healthy one, which means the differences between samples from infected ones are larger.

4.(b) T-test (two unpaired samples)

Choose one sided t test with 95% confidence interval, var.equal=TRUE

step 1: Hypothesis

$H_0 : \mu_1 \geq \mu_2$
$H_A : \mu_1 < \mu_2$

Parameter:

μ_1: is mean stem volume of 2-year-old seedlings propagated from virus-infected buds

μ_2 : is the mean stem volume of 2-year-old seedlings propagated from healthy buds

step 2: Test Statistics

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - E_0(\overline{Y}_1 - \overline{Y}_2)}{\sqrt{Var(\overline{Y}_1 - \overline{Y}_2)}}$$

$$t = \frac{\overline{Y}_1 - \overline{Y}_2 - 0}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim T_{n_1+n_2-2}$$

By calculations, $t = 7.01$

step 3: Find the p-value

p=3.222808e-07 , which is less than 0.05, so we should reject the H_0 . Thus, there is evidence that the mean stem volume of 2-year-old seedlings propagated from virus-infected buds is smaller than those propagated from healthy buds.

4.(c) Confidence Interval (two unpaired samples)

Choose two-sided t test with 95% confidence interval, var.equal=TRUE

$$P(\left|T_{n_1+n_2-2}\right| \geq t_{n_1+n_2-2,\alpha/2}) = \alpha$$

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{n_1+n_2-2,\alpha/2} \times \sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}$$

By calculations, 95% confidence interval for the difference of the mean stem volume of 2-year-old seedlings between the two groups is [654.3586,1206.7842], which means there are 95% probability that difference of the mean stem volume of 2-year-old seedlings between the two groups is in [654.3586,1206.7842].

4.(d) T-test (two unpaired samples)

Choose two-sided t test with 95% confidence interval, var.equal=TRUE

step 1: Hypothesis

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$

Parameters:

μ_1 : is the mean stem volume of 2-year-old seedlings propagated from virus-infected buds

μ_2 : is the mean stem volume of 2-year-old seedlings propagated from healthy buds

step 2: Test Statistics

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - E_0(\overline{Y}_1 - \overline{Y}_2)}{\sqrt{Var(\overline{Y}_1 - \overline{Y}_2)}}$$

$$t = \frac{\overline{Y}_1 - \overline{Y}_2 - 0}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim T_{n_1+n_2-2}$$

By r code, $t = 7.01$

step 3: Find the p-value

p=6.445616e-07 , which is less than 0.05, so we should reject the H_0 . Thus, there is evidence that the mean stem volume of 2-year-old seedlings propagated from virus-infected buds is different from those propagated from healthy buds.

4.(e) Assumptions

For(b)(c)(d), they are all unpaired two samples, so their assumptions should be the same as follows:

1) Independence within each sample and independence between two samples

2) Normality for each sample

3) Equal variance $\sigma_X^2 = \sigma_Y^2$

Assess the assumptions:

1) Independence

Considering how the data were collected, I suppose that they are selected randomly from two kinds of buds, they are more likely to be independent within the sample and with each other. But if not, they will not be independent.

2) Normality

Histogram: From the histogram plot, we can not clearly know that two samples are from normal distribution since the sample size is too small. So we use the more reliable tool— Q-Q plot.

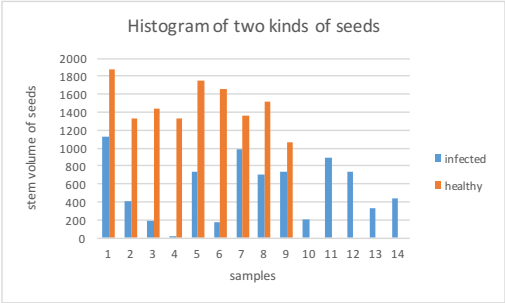


Figure 4.3 Histogram of buds

Q-Q Plot: From the plots, we can easily conclude that the random samples of both buds are from the normal population distribution. So the assumption is reasonable.

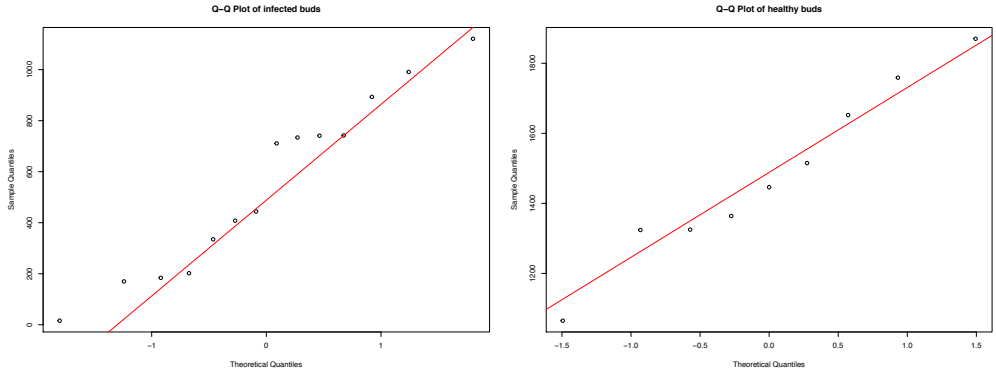


Figure 4.4 Q-Q Plot of buds

3) Equal variance

Boxplot: According to the graph, we can conclude that so the variance of both population is the same. So the assumption $\sigma_X^2 = \sigma_Y^2$ is reasonable.

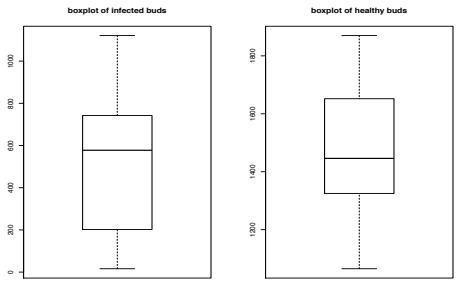


Figure 4.5 Box Plot of buds

Levene's test: According to the test, we can also found that p=0.1112>0.05, so the variance of both population is the same. So the assumption $\sigma_X^2 = \sigma_Y^2$ is reasonable.

```
> y=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444,1870,1324,1446,1325,1759,1652,1364,1515,1065)
> group=as.factor(c(1,1,1,1,1,1,1,1,1,1,1,1,1,   2,2,2,2,2,2,2,2,2))
> leveneTest(y = y, group=group)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  2.7656 0.1112
      21
```

Figure 4.6 Levene's test of buds

Remedial measures:

To make the two samples independent, we can choose two kinds of buds from several different locations, each location we randomly choose one healthy bud sample and one infected bud sample. Also, to solve the small size problem, we can do randomization test.

4.(f) Welch's T test (two unpaired samples)

Choose two-sided t-test with 95% confidence interval

Assumptions: Random Samples are not variance equal

step 1: Hypothesis

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$

Parameters:

μ_1: is the mean stem volume of 2-year-old seedlings propagated from virus-infected buds

μ_2: is the mean stem volume of 2-year-old seedlings propagated from healthy buds

step 2: Test Statistics

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - E_0(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{Var(\bar{Y}_1 - \bar{Y}_2)}}$$

$$t = \frac{\bar{Y}_1 - \bar{Y}_2 - 0}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim T_{n_1+n_2-2}$$

$t = 7.52$

step 3: Find the p-value

p=2.487409e-07 , which is less than 0.05, so we should reject the H_0 . Thus, there is evidence that the mean stem volume of 2-year-old seedlings propagated from virus-infected buds is different from those propagated from healthy buds.

step 4: Confidence Interval (two sided t test)

95% confidence interval for the difference of the mean stem volume of 2-year-old seedlings between the two groups is [672.9032,1188.2396], which means that there are 95% probability the difference of the mean stem volume of 2-year-old seedlings between the two groups is in [672.9032,1188.2396]

4.(g) Randomization test

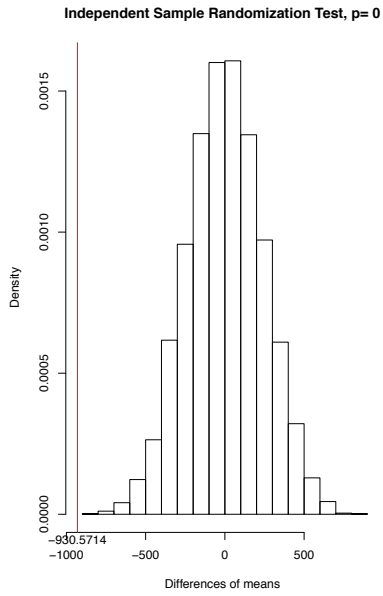


Figure 4.7 randomization test

From the results of randomization test, p value is 0, then we can conclude that we should reject the H_0 . Thus, there is evidence that the mean stem volume of 2-year-old seedlings propagated from virus-infected buds is different from those propagated from healthy buds.

4.(h) Nonparametric test

```
> # Wilcoxon Test (Nonparametric) ##################################################
> # Independent Two Samples (Wilcoxon rank sum test): "2-year-old seedlings" data
> x=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444)
> y=c(1870,1324,1446,1325,1759,1652,1364,1515,1065)
> wilcox.test(x, y)

        Wilcoxon rank sum test

data:  x and y
W = 1, p-value = 4.895e-06
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(x, y, exact=TRUE)

        Wilcoxon rank sum test

data:  x and y
W = 1, p-value = 4.895e-06
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(x, y, exact=FALSE)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 1, p-value = 0.000107
alternative hypothesis: true location shift is not equal to 0
```

Figure 4.8 nonparametric test

From the results of nonparametric test, p value is 0.000107, then we can conclude that we should reject the H_0. Thus, there is evidence that the mean stem volume of 2-year-old seedlings propagated from virus-infected buds is different from those propagated from healthy buds.

4.(i) Compare the results

Compare the results from T test, Welch's T test, randomization test and nonparametric test, we can all draw the conclusion that there is evidence that the mean stem volume of 2-year-old seedlings propagated from virus-infected buds is different from those propagated from healthy buds. And the samples are random.

4.(j) Sample sizes of n1 and n2

For unpaired two samples, we can use the following functions to calculate the values of n1 and n2.

$$n_1 = \frac{(\sigma_1^2 + \sigma_2^2 / k)(z_{1-\alpha/2} + z_{1-\beta})^2}{|\mu_1 - \mu_2|^2}, \quad k = n_2 / n_1$$

$$n_2 = \frac{(k\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2}{|\mu_1 - \mu_2|^2}$$

Use sample variance to replace the population variance.

> n1=1.430413

> n2=2.225087

5.(a) Scatterplot numerical statistics to summarize the data

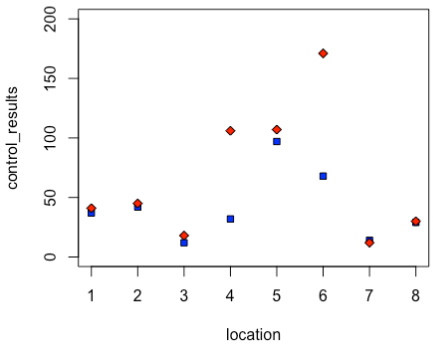


Figure 5.1 scatterplot of biological control and chemical control

	Biological control	Chemical control
median	34.5	43
mean	41.375	66.25
standard deviation	28.47524	55.91767

Table 5.2 numerical statistics of biological control and chemical control

Conclusions: The mean of the chemical control is high than that of biological control, which means it can capture more moths on average. But the standard of the chemical control is also larger than biological control's, which means it varies largely from plot to plot.

5.(b) T-test (one sample)

Choose two-sided t test with 95% confidence interval

step 1: Hypothesis

$H_0 : \mu = \mu_0$

$H_A : \mu \neq \mu_0$

Parameters:

μ : the mean of the difference of the two types of control

μ_0 : equals to 0

step 2: Test Statistics

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim T_{n-1}$$

the standard error is $s/\sqrt{n} = 40.18/\sqrt{8} = 14.21$

sample mean: $\bar{x} = -24.875$

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} = \text{-}1.75$$

step 3: Find the p-value

$2*P(T \geq 1.75)$ is about 0.1234, so at the 5% level, we should accept the H_0 . Thus, there is strong evidence that the means between the two types of control has no difference.

5.(c) Confidence Interval (two sided)

95% confidence interval for the difference of the means between the two types of control is [-58.4661,8.7161], which means there are 95% probability that the difference of the means between two types of control is in [-58.4661,8.7161]

5.(d) Assumptions

1) Independence within the sample

2) Normality for the sample

3) Outliers: can be sensitive

Assess the assumptions:

1) Independence

Considering how the data were collected, then we can judge whether it is independent or not. From the statement, it said that in the experiment, 8 plots were identified in a large field of alfalfa, and each plot was divided into two equal subplots. Within each plot, one of the two subplots was randomly assigned to be treated with the biological control; the other subplot was assigned a standard chemical treatment. l suppose that they are selected randomly from two kinds of buds, so they are more likely to be independent within the sample.

2) Normality

Histogram: From the histogram plot, we can not clearly know whether two samples are from normal distribution or not since the sample size is too small. So we use the more reliable tool— Q-Q plot.

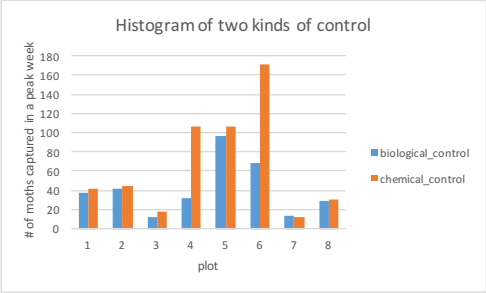


Figure 5.3 Histogram of two kinds of control

Q-Q Plot: From the plot, we can easily conclude that the difference of two kinds of control is not from the normal population distribution.

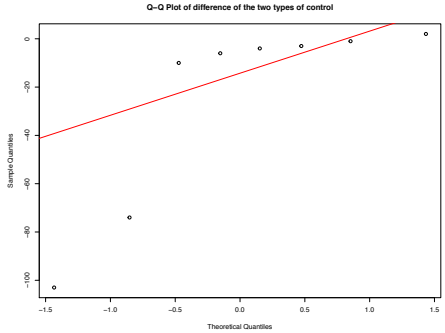


Figure 5.4 Q-Q Plot of difference of two kinds of control

3) Outliers

Boxplot: From the plot, we can fond that there is outlier in the sample, which t test will be sensitive to.

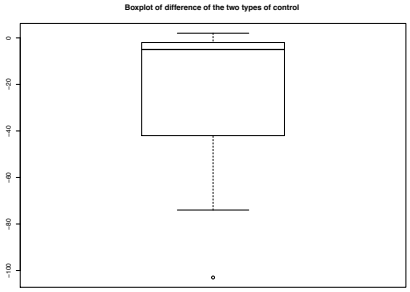


Figure 5.5 Box Plot of difference of two kinds of control

Remedial measures:

To make the assumption that samples are from the normal population distribution reasonable, we should do some log transformation to the data and make them be the normal distribution.

5.(e) T-test (One sample)

1）Transformation the data

To make the assumption that samples are from the normal population distribution reasonable, we should do transformation to the data. I transform them by log(data). To verify that they are from the normal population distribution, I do the QQ-plot of the transformed data again.

It can be found from the figure that the transformed data are from normal distribution. Then we can do t-test for them.

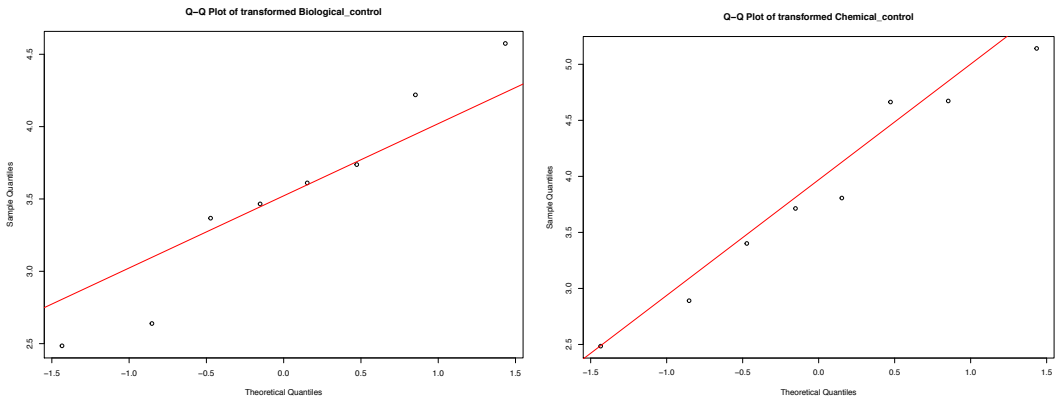


Figure 5.6 QQ-plot of the transformed data

2) T-test

Choose two-sided t test with 95% confidence interval

step 1: Hypothesis

$H_0 : \mu = \mu_0$

$H_A : \mu \neq \mu_0$

Parameters:

μ : the mean of the difference of the two types of control

μ_0 : equals to 0

step 2: Test Statistics

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim T_{n-1}$$

$$t = \frac{\overline{x} - 0}{s/\sqrt{n}} = -1.975$$

step 3: Find the p-value

$2*P(T \geq -1.975)$ is about 0.0888, so at the 5% level, we should accept the H_0. Thus, there is no evidence that the means are different between the two types of control.

step 4: Confidence Interval

$$P(-t_{n-1,\frac{\alpha}{2}} \leq T_{n-1} \leq t_{n-1,\frac{\alpha}{2}}) = 1-\alpha$$

$$\mu_X \in [\overline{x} - t_{n-1,\frac{\alpha}{2}} \times (S/\sqrt{n}) \leq T_{n-1} \leq \overline{x} + t_{n-1,\frac{\alpha}{2}} \times (S/\sqrt{n})]$$

$$\mu_X \in [-0.7346, 0.0659]$$

95% confidence interval for the difference of the means between the two types of control is [-0.7346,0.0659], which means there are 95% probability that the difference of the means between two types of control is in [-0.7346,0.0659]

5.(f) Randomization test

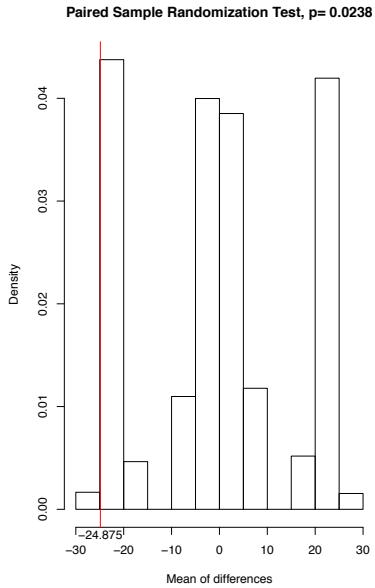


Figure 5.6 randomization test

From the results of randomization test, p value is 0.0238<0.05, then we can conclude that we should reject the H_0. Thus, there is strong evidence that means are different between the two types of control.

5.(g) Nonparametric test

```
> # Paired Sample (Wilcoxon signed rank test): "biological control and chemical control" data
> biological_control <- c(37,42,12,32,97,68,14,29)
> chemical_control <- c(41,45,18,106,107,171,12,30)
> wilcox.test(biological_control, chemical_control, paired=TRUE)

        Wilcoxon signed rank test

data:  biological_control and chemical_control
V = 2, p-value = 0.02344
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(biological_control, chemical_control, paired=TRUE, exact=TRUE)

        Wilcoxon signed rank test

data:  biological_control and chemical_control
V = 2, p-value = 0.02344
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(biological_control, chemical_control, paired=TRUE, exact=FALSE)

        Wilcoxon signed rank test with continuity correction

data:  biological_control and chemical_control
V = 2, p-value = 0.02997
alternative hypothesis: true location shift is not equal to 0
```

Figure 5.7 nonparametric test

From the results of nonparametric test, p value is 0.02997<0.05, then we can conclude that we should reject the H_0. Thus, there is strong evidence that means are different between the two types of control.

5.(h) Compare the results

Compare all the results from two T test, we can draw the conclusion that there is evidence that the means between the two types of control is no difference. But for randomization test and nonparametric test, it indicates that there is evidence that the means between the two types of control are different, and l think that maybe because it looses the boundary.

6. A cool product（group work is in the appendix）

Our product is a report to test a common sense that the height of male is usually larger than that of female.

Our sample is from google (http://ncdrisc.org/data-downloads-height.html) and both paired male and female's height from 200 countries in 1996. The sample is from normal distribution.

step 1: Hypothesis

Assumptions:

1) two samples (the height of male and female) are paired and independent, it is reasonable because the height of male and female are one-to-one and randomly selected in each country.

2) $X_i \overset{iid}{\sim} N(\mu, \frac{\sigma^2}{n})$, σ^2 is unknown. It is reasonable because of the **central limit theorem** that when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed.

Hypothesis:

$H_0 : \mu = 0$ represents that there is no difference between the mean of the height of male and female.

$H_A : \mu > 0$ represents that the mean of male's height is larger than that of female's.

μ is the mean of the difference between the height of male and female worldwide in 1996.

step 2: Test Statistics (one sample two sided t test with 95%CI)

We choose t-test and the test statistics is $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim T_{n-1}$ since sample variables are from normal distribution population.

We let $\alpha = 0.05$, and then with r code, we get t =86.83013

step 3: Find the p-value

With r code, we can also get the p-value=1.867356e-160 , which is less than $\alpha = 0.05$, so we reject H_0 and accept H_A. In this way, there is strong evidence that the mean of male's height is larger than that of female's.

step 4: Confidence Interval (two sided with 95% confidence level)

$$P(-t_{n-1,\frac{\alpha}{2}} \le T_{n-1} \le t_{n-1,\frac{\alpha}{2}}) = 1-\alpha \quad \Rightarrow \quad \mu_X \in [\bar{x} - t_{n-1,\frac{\alpha}{2}} \times (S/\sqrt{n}) \le T_{n-1} \le \bar{x} + t_{n-1,\frac{\alpha}{2}} \times (S/\sqrt{n})]$$

With r code, we get $\mu_X \in [11.52, 12.06]$ which means that there is 95% probability that the difference of the means of height between the male and female is in[11.52, 12.06]

Conclusions:

Our product can easily explain the three concepts. By the whole report, it explains how **hypothesis testing** works to test whether the common sense that the height of male is usually larger than that of female is true and the answer is yes. For **confidence interval**, it gives a 95% probability that male is higher than women about 11.52 to 12.06 centimeters on average. And the **central limit theorem** makes the hypothesis testing's assumption reasonable.

Appendix

```
2(a)(b)(c)
# sample mean
S = 100
par(mfrow = c(1,3)) # plot 4 figures
for(n in c(10,40,160)){
   sample_mean=rep(NA,S)
   for(i in 1:S){
      sample_norm=rnorm(n,2,2)
      sample_mean[i]=mean(sample_norm)
   }
   if (n == 10){
      hist(sample_mean, xlim=c(0,4), ylim=c(0,30),xlab="sample mean",
            col=rgb(0,0,1,1/4), main="random sample(n=10) from N(2,4)",
            breaks = seq(0,4,0.2))
   }
   if (n == 40){
      hist(sample_mean, xlim=c(0,4), ylim=c(0,30),xlab="sample mean",
            col=rgb(0,1,1,1/4), main="random sample(n=40) from N(2,4)",
            breaks = seq(0,4,0.2))
   }
   if (n == 160){
      hist(sample_mean, xlim=c(0,4), ylim=c(0,30),xlab="sample mean",
            col=rgb(0,1/4,1,1/4), main="random sample(n=160) from
N(2,4)",
            breaks = seq(0,4,0.2))
   }
}
# sample variance
S = 100
par(mfrow = c(1,3)) # plot 4 figures
for(n in c(10,40,160)){
   sample_var=rep(NA,S)
   for(i in 1:S){
      sample_norm=rnorm(n,2,2)
      sample_var[i]=var(sample_norm)
   }
   if (n == 10){
      hist(sample_var, xlim=c(0,10), ylim=c(0,100),xlab="sample var",
            col=rgb(0,0,1,1/4), main="random sample(n=10) from N(2,4)",
            breaks = seq(0,10,0.5))
   }
```

```
    if (n == 40){
      hist(sample_var, xlim=c(0,10), ylim=c(0,100),xlab="sample var",
           col=rgb(0,1,1,1/4), main="random sample(n=40) from N(2,4)",
           breaks = seq(0,10,0.5))
    }
    if (n == 160){
      hist(sample_var, xlim=c(0,10), ylim=c(0,100),xlab="sample var",
           col=rgb(0,1/4,1,1/4), main="random sample(n=160) from
N(2,4)",
           breaks = seq(0,10,0.5))
    }
}
```

2(e)

```
# sample mean
S = 100
par(mfrow = c(1,3)) # plot 4 figures
for(n in c(10,40,160)){
   sample_mean=rep(NA,S)
   for(i in 1:S){
      sample_norm=rbinom(n,10,0.2)
      sample_mean[i]=mean(sample_norm)
   }
   if (n == 10){
      hist(sample_mean, xlim=c(0,5), ylim=c(0,50),xlab="sample mean",
           col=rgb(0,0,1,1/4), main="random sample(n=10) from
B(10,0.2)",
           breaks = seq(0,5,0.5))
   }
   if (n == 40){
      hist(sample_mean, xlim=c(0,5), ylim=c(0,50),xlab="sample mean",
           col=rgb(0,1,1,1/4), main="random sample(n=40) from
B(10,0.2)",
           breaks = seq(0,5,0.5))
   }
   if (n == 160){
      hist(sample_mean, xlim=c(0,5), ylim=c(0,50),xlab="sample mean",
           col=rgb(0,1/4,1,1/4), main="random sample(n=160) from
B(10,0.2)",
           breaks = seq(0,5,0.5))
   }
}
# sample variance
S = 100
par(mfrow = c(1,3)) # plot 4 figures
```

```
for(n in c(10,40,160)){
   sample_var=rep(NA,S)
   for(i in 1:S){
      sample_norm=rbinom(n,10,0.2)
      sample_var[i]=var(sample_norm)
   }
   if (n == 10){
      hist(sample_var, xlim=c(0,10), ylim=c(0,100),xlab="sample var",
            col=rgb(0,0,1,1/4), main="random sample(n=10) from
B(10,0.2)",
            breaks = seq(0,10,0.5))
   }
   if (n == 40){
      hist(sample_var, xlim=c(0,10), ylim=c(0,100),xlab="sample var",
            col=rgb(0,1,1,1/4), main="random sample(n=40) from
B(10,0.2)",
            breaks = seq(0,10,0.5))
   }
   if (n == 160){
      hist(sample_var, xlim=c(0,10), ylim=c(0,100),xlab="sample var",
            col=rgb(0,1/4,1,1/4), main="random sample(n=160) from
B(10,0.2)",
            breaks = seq(0,10,0.5))
   }
}
```

3(b)(c)(d)

```
# outlier
par(mfrow=c(1,2))
subject<-c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20)
Pretest<-c(30,28,31,26,20,30,34,15,28,20,30,29,31,29,34,20,26,25,31,29)
boxplot(Pretest,range = 1.5,main="outlier of Pretest")
Posttest<-c(29,30,32,30,16,25,31,18,33,25,32,28,34,32,32,27,28,29,32,32)
boxplot(Posttest,range = 1.5,main="outlier of Posttest")
# skewness
par(mfrow=c(1,2))
Pretest<-c(30,28,31,26,20,30,34,15,28,20,30,29,31,29,34,20,26,25,31,29)
hist(Pretest,main="skewness of Pretest")
Posttest<-c(29,30,32,30,16,25,31,18,33,25,32,28,34,32,32,27,28,29,32,32)
hist(Posttest,main="skewness of Posttest")
# t-test
out = t.test(x=c(-1,2,1,4,-4,-5,-3,3,5,5,2,-1,3,3,-2,7,2,4,1,3),alternative =
"greater", mu = 0, conf.level = .95)
out$statistic
out$p.value
```

```
# confidence interval
out = t.test(x=c(-1,2,1,4,-4,-5,-3,3,5,5,2,-1,3,3,-2,7,2,4,1,3),alternative =
"two.sided", mu = 0, conf.level = .90)
out$conf.int
```

3.(e)

```
# Find minimum n
for(n in 1:2000){
   prob=pnorm(qnorm(0.975,0,1)/sqrt(n),2,1/n)-
pnorm(qnorm(0.025,0,1)/sqrt(n),2,1/n)
   if(prob <= 0.2){
      break
   }
   print(n)
}
```

4(a)(b)(c)

```
# summary plots
Healthy<-c(1870,1324,1446,1325,1759,1652,1364,1515,1065)
Infected<-c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444)
median(Healthy)
mean(Healthy)
sd(Healthy)
median(Infected)
mean(Infected)
sd(Infected)
boxplot(Healthy,main='Healthy')
boxplot(Infected,main='Infected')
# t-test
x=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444)
y=c(1870,1324,1446,1325,1759,1652,1364,1515,1065)
out = t.test(x, y , alternative = "less", mu = 0, var.equal=TRUE, paired =
FALSE, conf.level = .95)
out$statistic
out$p.value
# confidence interval
x=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444)
y=c(1870,1324,1446,1325,1759,1652,1364,1515,1065)
out = t.test(y, x, alternative = "two.sided", mu = 0, paired = FALSE,
var.equal=TRUE, conf.level = .95)
out$conf.int
```

4(d)

```
# t-test
x=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444)
y=c(1870,1324,1446,1325,1759,1652,1364,1515,1065)
out = t.test(x, y , alternative = "two.sided", mu = 0, var.equal=TRUE,paired
= FALSE,conf.level = .95)
out$statistic
out$p.value
```

4(e)

```
# boxplot
x=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444)
y=c(1870,1324,1446,1325,1759,1652,1364,1515,1065)
par(mfrow=c(1,2))
boxplot(x,main="boxplot of infected buds")
boxplot(y,main="boxplot of healthy buds")
# qq-plot
x=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444)
qqnorm(x,main = "Q-Q Plot of infected buds")
qqline(x, col=2, lwd=2)
y=c(1870,1324,1446,1325,1759,1652,1364,1515,1065)
qqnorm(y,main = "Q-Q Plot of healthy buds")
qqline(y, col=2, lwd=2)
# Levene's test
x=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444)
y=c(1870,1324,1446,1325,1759,1652,1364,1515,1065)
y=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444,1870,1324,
1446,1325,1759,1652,1364,1515,1065)
group=as.factor(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,      2,2,2,2,2,2,2,2,2))
leveneTest(y = y, group=group)
```

4(f)

```
# Welch's T test
x=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444)
y=c(1870,1324,1446,1325,1759,1652,1364,1515,1065)
out = t.test(y,x , alternative = "two.sided", mu = 0, paired =
FALSE,conf.level = .95)
out$statistic
out$p.value
out$conf.int
```

4(g)(h)

```
# Randomization Test
x=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444)
y=c(1870,1324,1446,1325,1759,1652,1364,1515,1065)
set.seed(18);
rand.test(x, y, paired=F)
# Wilcoxon Test
x=c(1121,408,184,16,741,170,991,711,734,202,893,742,335,444)
y=c(1870,1324,1446,1325,1759,1652,1364,1515,1065)
wilcox.test(x, y)
wilcox.test(x, y, exact=TRUE)
wilcox.test(x, y, exact=FALSE)
```

5(a)

```
# scatterplot
par(mfrow=c(1,1))
location<-c(1,2,3,4,5,6,7,8)
Biological_control<-c(37,42,12,32,97,68,14,29)
Chemical_control<-c(41,45,18,106,107,171,12,30)
plot(location,Biological_control,ylim=c(0,200),pch=22,bg="blue",xlab =
"location", ylab = "control_results")
par(new=T)
plot(location,Chemical_control,ylim=c(0,200),pch=23,bg="red",axes =
FALSE,xlab = "", ylab = "")
# statistics and boxplot
par(mfrow=c(1,2))
Biological_control<-c(37,42,12,32,97,68,14,29)
Chemical_control<-c(41,45,18,106,107,171,12,30)
median(Biological_control)
mean(Biological_control)
sd(Biological_control)
median(Chemical_control)
mean(Chemical_control)
sd(Chemical_control)
boxplot(Biological_control,main='Biological_control')
boxplot(Chemical_control,main='Chemical_control')
```

5(b)

```
# t-test and confidence interval
biological_control <- c(37,42,12,32,97,68,14,29)
chemical_control <- c(41,45,18,106,107,171,12,30)
x=c(biological_control-chemical_control)
out = t.test(x,alternative = "two.sided", mu = 0,conf.level = .95)
out$statistic
out$p.value
out$conf.int
```

5(d)

```
# qq-plot
biological_control <- c(37,42,12,32,97,68,14,29)
chemical_control <- c(41,45,18,106,107,171,12,30)
x=c(biological_control-chemical_control)
qqnorm(x,main = "Q-Q Plot of difference of the two types of control")
qqline(x, col=2, lwd=2)
# boxplot
biological_control <- c(37,42,12,32,97,68,14,29)
chemical_control <- c(41,45,18,106,107,171,12,30)
x=c(biological_control-chemical_control)
boxplot(x,main="Boxplot of difference of the two types of control ")
```

5(e)

```
# qq-plot
par(mfrow=c(1,1))
Biological_control<-c(37,42,12,32,97,68,14,29)
Chemical_control<-c(41,45,18,106,107,171,12,30)
x=log(Biological_control)
qqnorm(x,main = "Q-Q Plot of transformed Biological_control")
qqline(x, col=2, lwd=2)
# qq-plot
Biological_control<-c(37,42,12,32,97,68,14,29)
Chemical_control<-c(41,45,18,106,107,171,12,30)
y=log(Chemical_control)
qqnorm(y,main = "Q-Q Plot of transformed Chemical_control")
qqline(y, col=2, lwd=1)
#t-test
z=x-y
out = t.test(z,alternative = "two.sided", mu = 0, var.equal=TRUE,conf.level
= .90)
out$statistic
out$p.value
out$conf.int
```

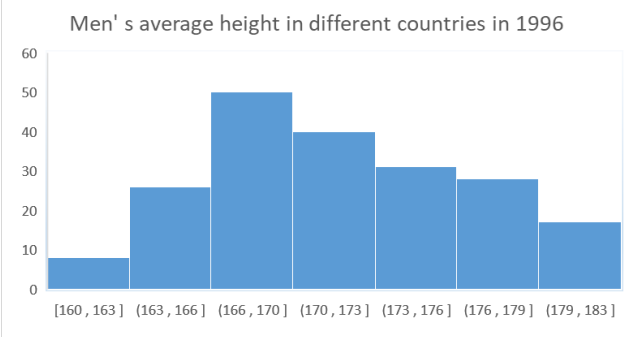
5(f)(g)

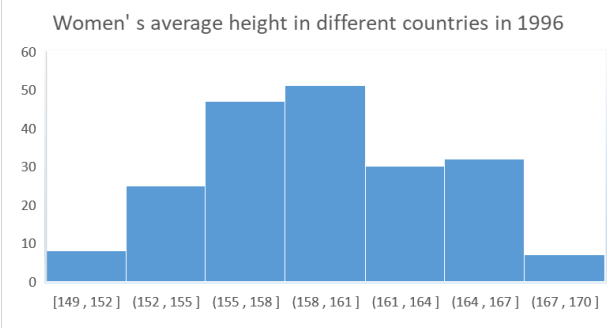
```
# Randomization Test
biological_control <- c(37,42,12,32,97,68,14,29)
chemical_control <- c(41,45,18,106,107,171,12,30)
set.seed(18);
rand.test(biological_control,chemical_control, paired = T)
# Wilcoxon Test
biological_control <- c(37,42,12,32,97,68,14,29)
chemical_control <- c(41,45,18,106,107,171,12,30)
wilcox.test(biological_control, chemical_control, paired=TRUE)
```

```
wilcox.test(biological_control, chemical_control, paired=TRUE,
exact=TRUE)
wilcox.test(biological_control, chemical_control, paired=TRUE,
exact=FALSE)
```

6

```
# t-test
a<-read.csv("height worldwide 1996.csv",header = T)
x=c(a$difference)
out = t.test(x, alternative = "greater", mu = 0, var.equal=TRUE,conf.level
= .95)
out$statistic
out$p.value
out = t.test(x, alternative = "two.sided", mu = 0, var.equal=TRUE,conf.level
= .95)
out$conf.int
# histogram of samples
```





Group work

1.Naiqing Cai, Statistics (former major in Financial Engineering)

Pick a dataset

Transform the data to an appropriate form

Do t-test and confidence interval

2.Hao Pan, Statistics (former major in Finance)

Analyze the data whether fits the assumptions

Do graphs

Do t-test and confidence interval