**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.

Rplot01.pdfRplot02.pdf

Figure 2.1 sample mean Figure 2.2 sample variance

**2.(d)**

From the histograms in (a)–(c), l 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.

2(e)mean.pdf2(e)var.pdf

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)**







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.

is the variance of sample mean, and  is the variance of samples of the population.

**2.(g)**









is the variance of sample mean, and is the variance of samples of the population.

**3.(a) Hypothesis**



Parameters:

: where 109-11.png is the mean of the population posttest scores minus the mean of the population pretest scores.

: equals to 0

**3.(b) Graphical check**

3(b)outlier.pdf3(b)skewness.pdf

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**



**step 2: Test Statistics**



the standard error is 

sample mean: 



**step 3: Find the p-value**

 is 0.0286, less than 0.05, so at the 5% level, so we should reject the  . 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







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**







From r code, we can find that 

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

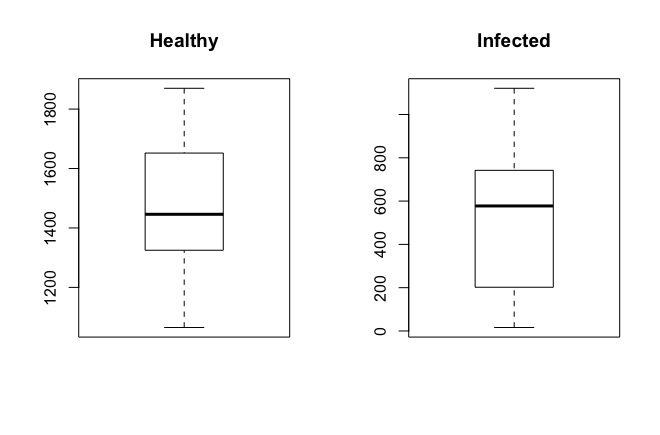


Figure 4.1 boxplot of healthy and infected buds

**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**



Parameter:

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

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

**step 2: Test Statistics**





By calculations, 

**step 3: Find the p-value**

, which is less than 0.05, so we should reject the . 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





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**



Parameters:

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

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

**step 2: Test Statistics**





By r code, 

**step 3: Find the p-value**

, which is less than 0.05, so we should reject the . 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 

**Assess the assumptions:**

1) Independence

Considering how the data were collected, l 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.

../qqplot1.pdf../qqplot2.pdf

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  is reasonable.

4boxplot.pdf

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  is reasonable.

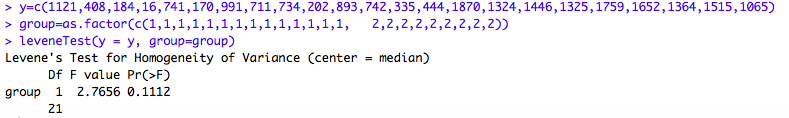


Figure 4.6 Levene’s test of buds

**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**



Parameters:

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

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

**step 2: Test Statistics**







**step 3: Find the p-value**

, which is less than 0.05, so we should reject the . 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**

Rplot.pdf

Figure 4.7 randomization test

From the results of randomization test, p value is 0, then we can conclude that we should reject the . 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**

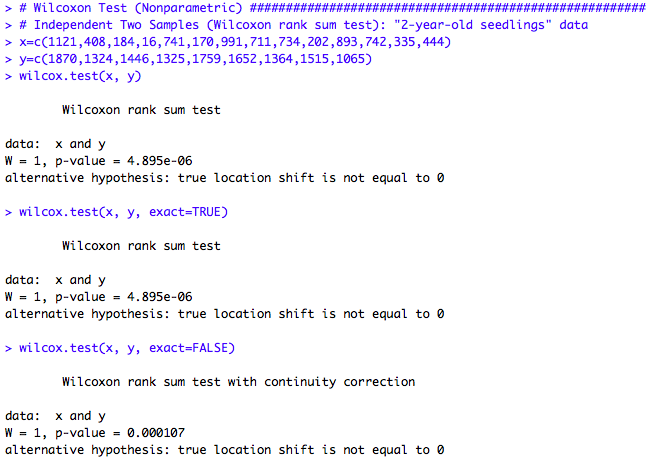


Figure 4.8 nonparametric test

From the results of nonparametric test, p value is 0.000107, then we can conclude that we should reject the . 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.

, 



Use sample variance to replace the population variance.

> n1=1.430413

> n2=2.225087

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

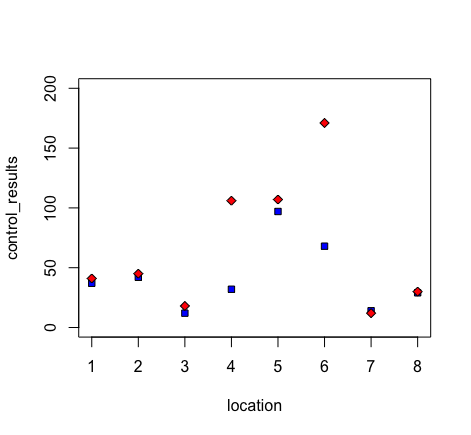


Figure 5.1 scatterplot of biological control and chemical control

|  |  |  |
| --- | --- | --- |
|  | 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**



Parameters:

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

: equals to 0

**step 2: Test Statistics**



the standard error is 

sample mean: 



**step 3: Find the p-value**

 is about 0.1234, so at the 5% level, we should accept the  . 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.

601/hw1/qqplot3.pdf

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.

5boxplot.pdf

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.

qqplot%20trans1.pdfqqplot%20trans2.pdf

Figure 5.6 QQ-plot of the transformed data

2) T-test

Choose two-sided t test with 95% confidence interval

**step 1: Hypothesis**



Parameters:

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

: equals to 0

**step 2: Test Statistics**





**step 3: Find the p-value**

 is about 0.0888, so at the 5% level, we should accept the  . Thus, there is no evidence that the means are different between the two types of control.

**step 4: Confidence Interval**



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**

5(f).pdf

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 . Thus, there is strong evidence that means are different between the two types of control.

**5.(g) Nonparametric test**

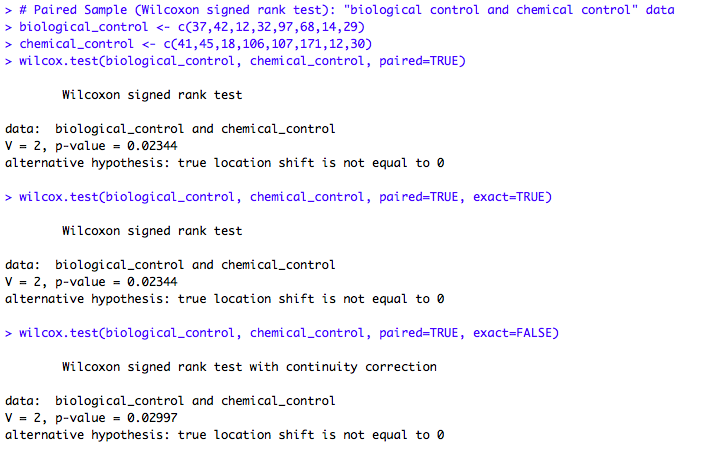


Figure 5.7 nonparametric test

From the results of nonparametric test, p value is 0.02997<0.05, then we can conclude that we should reject the . 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) ，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:

 represents that there is no difference between the mean of the height of male and female.

 represents that the mean of male’s height is larger than that of female’s.

 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  since sample variables are from normal distribution population.

We let , 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 , so we reject  and accept . 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)**

With r code, we get  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 |