Statistics 601 – Assignment 1 — Due September 20

Name: Naiqing Cai email:ncai5@wisc.edu

1. 1）graphical methods: Q–Q plot

**Definition**

It is a plot for comparing two probability distributions by plotting their quantiles against each other.

**Advantages**

It can not only test whether the sample is to specify the distribution, also can detect whether the two samples have the same distribution.

**Example (Two classes exam performance data)**

Many '+' number deviate from the reference line, so it can be considered that two class grades don't follow the same distribution.



1. 2）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)**

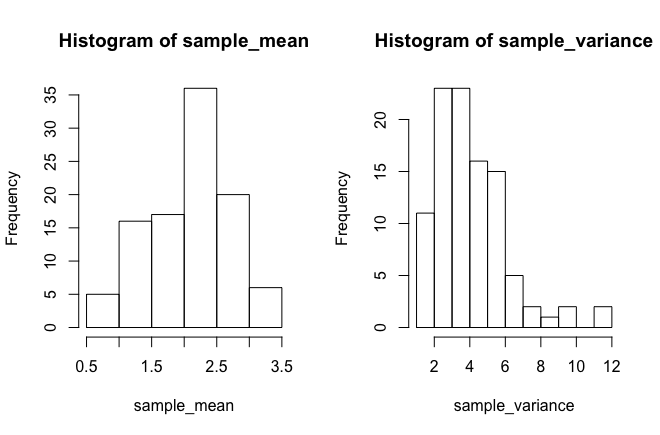
Celsius: [-10, 2, -8, 5, 7, -15] Fahrenheit: [13, 15, 6, 14, 1, 10]

The sample standard deviations are 9.02 and 5.42, respectively.

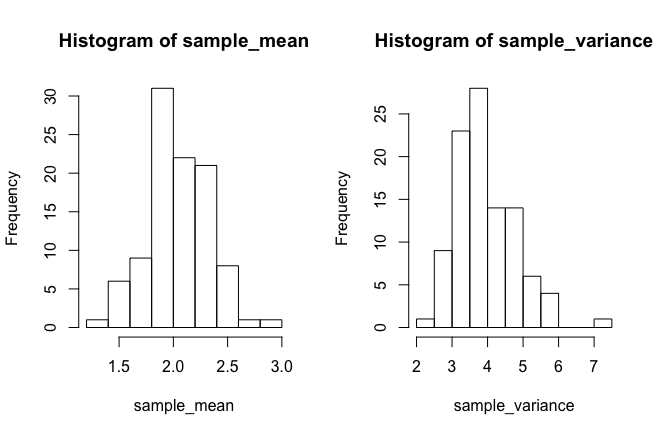
The CV of the first set is 9.02/-3.17 = -2.85. For the second set (which are the same temperatures) it is 5.42/9.83 = 0.55

2.

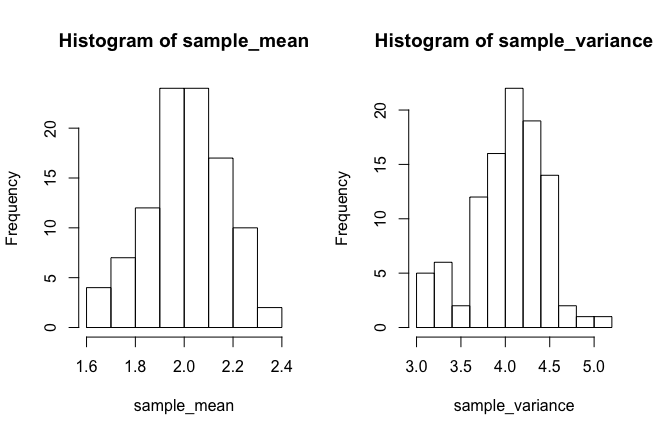
(a) n=10



(b) n=40



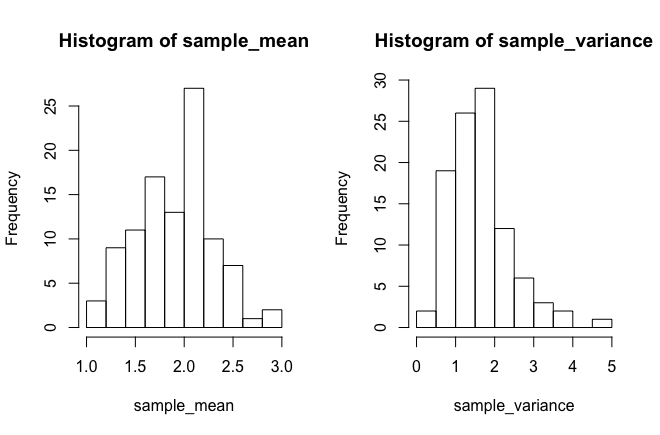
(c) n=160



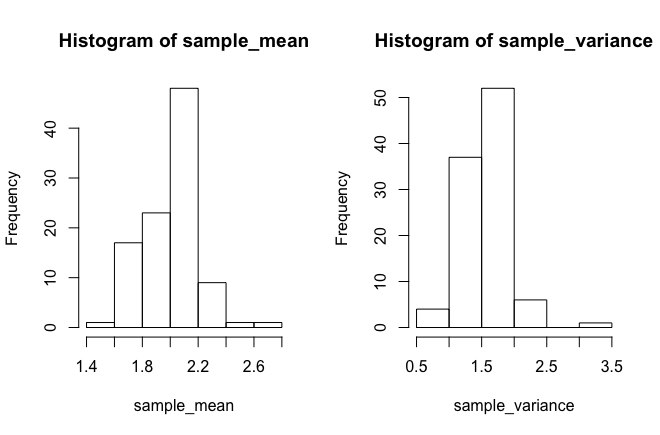
(d)

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. Also, when n=10 or n=40, which is relatively small, the histograms of sample variance are right-skewed, but when n=160, which means observations are bigger, the histograms are symmetric-shaped.

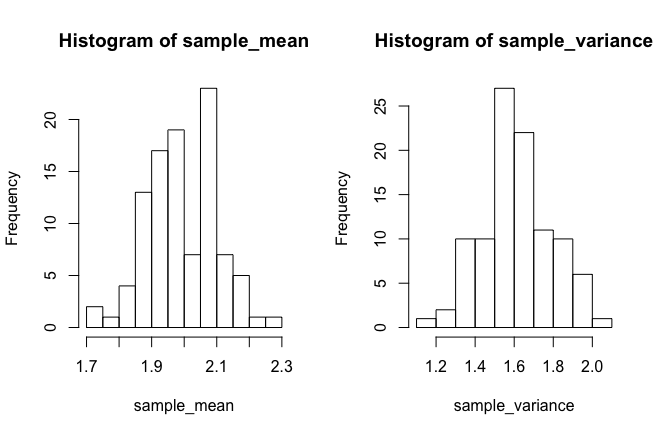
(e)



n=10



n=40



n=160

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. Also, when n=10 or n=40, which is relatively small, the histograms of sample variance are right-skewed, but when n=160, which means observations are bigger, the histograms of sample variance are symmetric-shaped.

(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.

(g)

3.

(a)

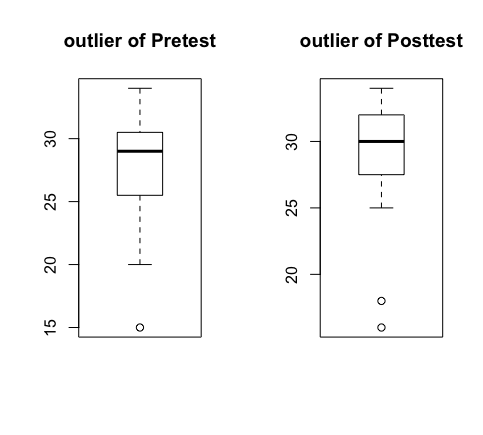
H0:

Ha :

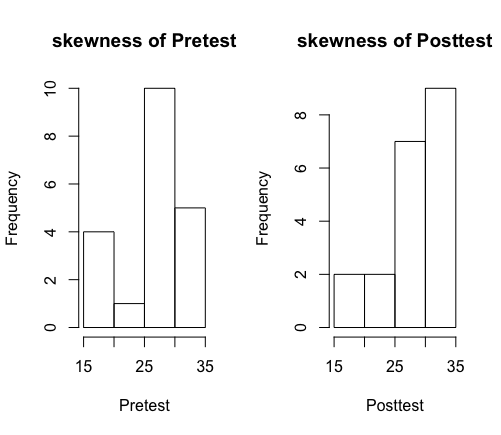
Parameter

(b)

outliers

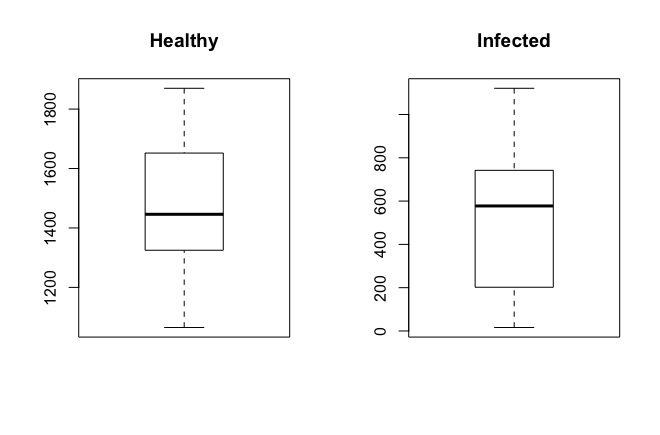


strong skewness



4.

(a)boxplot and numerical statistic to summarize the data

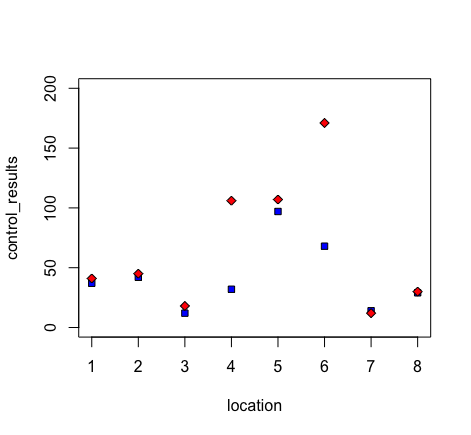


|  |  |  |
| --- | --- | --- |
|  | Healthy | Infected |
| median | 1446 | 577.5 |
| mean | 1480 | 549.4286 |
| standard deviation | 248.9849 | 343.4586 |

**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.

5.



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

**Appendix**

|  |
| --- |
| **2(a)** #n=10  sample\_mean = c()  sample\_variance = c()  for (i in 1:100)  { sample<-rnorm(10,2,2)  sample\_mean<-cbind(sample\_mean,mean(sample))  sample\_variance<-cbind(sample\_variance,var(sample))  }  par(mfrow=c(1,2))  hist(sample\_mean)  hist(sample\_variance) |
| **2(b)** #n=40  sample\_mean = c()  sample\_variance = c()  for (i in 1:100)  { sample<-rnorm(40,2,2)  sample\_mean<-cbind(sample\_mean,mean(sample))  sample\_variance<-cbind(sample\_variance,var(sample))  }  par(mfrow=c(1,2))  hist(sample\_mean)  hist(sample\_variance)  **2(c)** #n=160  sample\_mean = c()  sample\_variance = c()  for (i in 1:100)  { sample<-rnorm(160,2,2)  sample\_mean<-cbind(sample\_mean,mean(sample))  sample\_variance<-cbind(sample\_variance,var(sample))  }  par(mfrow=c(1,2))  hist(sample\_mean)  hist(sample\_variance) |
| **2(e)**  #n=10  sample\_mean = c()  sample\_variance = c()  for (i in 1:100)  { sample<-rbinom(10,10,0.2)  sample\_mean<-cbind(sample\_mean,mean(sample))  sample\_variance<-cbind(sample\_variance,var(sample)  )  }  par(mfrow=c(1,2))  hist(sample\_mean)  hist(sample\_variance)  #n=40  sample\_mean = c()  sample\_variance = c()  for (i in 1:100)  { sample<-rbinom(40,10,0.2)  sample\_mean<-cbind(sample\_mean,mean(sample))  sample\_variance<-cbind(sample\_variance,var(sample)  )  }  par(mfrow=c(1,2))  hist(sample\_mean)  hist(sample\_variance)  #n=160  sample\_mean = c()  sample\_variance = c()  for (i in 1:100)  { sample<-rbinom(160,10,0.2)  sample\_mean<-cbind(sample\_mean,mean(sample))  sample\_variance<-cbind(sample\_variance,var(sample)  )  }  par(mfrow=c(1,2))  hist(sample\_mean)  hist(sample\_variance) |
| **3(b)**  # 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") |
|  |
|  |