目录

[1. Input: 1](#_Toc370430681)

[2.Seq 1](#_Toc370430682)

[3.Graph 1](#_Toc370430683)

[4.Distribution: CDF-P, PDF-d 2](#_Toc370430684)

[5.Confidence Interval 3](#_Toc370430685)

[6.Hypothesis Testing 4](#_Toc370430686)

[7.Chi-square Test(Lecture 15) 4](#_Toc370430687)

[8.Test for Normality 5](#_Toc370430688)

[9.Not normal: 5](#_Toc370430689)

[10.Correlation 5](#_Toc370430690)

# 1. Input:

> site<-"http://people.sc.fsu.edu/~jburkardt/datasets/census/census\_2010.txt"

> data <- read.table(site, header=F)

> site<-"http://people.sc.fsu.edu/~jburkardt/datasets/stats/height\_female\_baby.csv"

> data <- read.table(site, header=T, sep=",")

>head(data, n=10)# print first 10 rows of the data

>tail(data, n=10) # print last 10 rows of the data

> Data=read.table("F:/xhtest.txt",header=T,na.strings="NA",sep="\t")

# 2.Seq

> seq(1,10, by=3) # We can simply use seq(1,10,3)

[1] 1 4 7 10

> seq(1,50,length=8)

[1] 1 8 15 22 29 36 43 50

> rep(1,6)

[1] 1 1 1 1 1 1

> rep(c(1,2,3,4), each=3)

[1] 1 1 1 2 2 2 3 3 3 4 4 4

> rep(c(1,2,3,4), 3)

[1] 1 2 3 4 1 2 3 4 1 2 3 4

# 3.Graph

>hist(x)

> Country=c("United Kingdom","Czech Republic","Italy","Germany","France","Canada","U.S.")

> Amount=c(1992,1106,2212,2808,2561,2792,4887)

> Health=data.frame(Country,Amount)

> names(Amount)=Country

>barplot(Amount,col=c("1","2","3","4","5","6","7"),main="spending on health care",xlab="Country",ylab="Amounnts",ylim=c(0,5000), names.arg=c("1","2","3","4","5","6","7"))

> x1=PlantGrowth$weight[ which(PlantGrowth$group=="ctrl")]

> x2=PlantGrowth$weight[ which(PlantGrowth$group=="trt1")]

> x3=PlantGrowth$weight[ which(PlantGrowth$group=="trt2")]

> boxplot(x1,x2,x3,main="Data of Plant Weight",ylab="weight",xlab="Condition")

> hist(Min.Price,50, freq=FALSE,main="Density Curve of Min.Price",col=2)

> lines(density(Min.Price,width=10,n=200))

>students<-c(120, 200, 350, 230,125)

>names(students)=c("Maths","Science","Engineering",

"Education","Technology")

>pie(students,col=c("1","2","3","4","5"),

main="Student Enrollment")

x<-seq(-5,5,0.1)

y1<-dnorm(x)

y2<-dcauchy(x)

y3<-0.5\*dexp(abs(x))

yrange<-range(y1,y2,y3)

plot(x,y1,xlab="x",ylab="f(x)",lty=1, type="l",xlim=c(-5,5),ylim=yrange,col=1)

par(new=TRUE)

plot(x,y2,xlab="",ylab="",lty=3,type="l",xlim=c(-5,5),ylim=yrange,col=2)

par(new=TRUE)

plot(x,y3,xlab="",ylab="",lty=2,type="l",xlim=c(-5,5),ylim=yrange,col=4)

legend(1,.5,legend=c("N(0,1)","C(0,1)","L(0,1)"),lty=c(1,3,2),col=c(1,2,4))

title(cex=1,"probability density functions of standard Normal, standard Cauchy and

\n standard Laplace distributions")

# 4.Distribution: CDF-P, PDF-d

> set.seed(20)

> rnorm(5)

[1] 1.1626853 -0.5859245 1.7854650 -1.3325937 -0.4465668

> set.seed(20)

> rnorm(5)

[1] 1.1626853 -0.5859245 1.7854650 -1.3325937 -0.4465668

curve(dnorm(x,0,0.3),from=-3,to=3,col="blue")

curve(dnorm(x,0,1),from=-3, to=3, col="red", add=T)

Obtain 95% quantile for the standard normal distribution

> qnorm(0.95)

[1] 1.644854

curve(df(x,2,4),from=0,to=10,col=1, ylim=c(0,1),xlim=c(0,10), lwd=2,

ylab="f(x)",xlab="x",main="pdf of F-distribution", lty=1)

curve(df(x,5,5),from=0, to=10, col=2, add=T,lty=2,lwd=2)

curve(df(x,10,15),from=0,to=10,col=3, add=T,lty=3,lwd=2)

curve(df(x,15,15),from=0, to=10, col=4, add=T,lty=4,lwd=2)

legend(4,0.9,legend=c(expression(n1==2~~n2==4),expression(n1==5~~n2==5),

expression(n1==10~~n2==15), expression(n1==15~~n2==15)),lty=1:4,lwd=2,

col=c(1,2,3,4))

# 5.Confidence Interval

Confidence interval

1) The water works commission wishes to know the mean household usage of water by the residents of a small town. Suppose a sample of 1403 families is drawn from the town with mean=18 and assume that the standard deviation is 1.7. Construct a99% confidence interval for the mean number of gallons of water.

> xbar=18

> n=1403

> Z=qnorm(0.995)

> sigma=1.7

> round(c(xbar-Z\*sigma/sqrt(n),xbar+Z\*sigma/sqrt(n)),2)

[1] 17.88 18.12

3) A research scholar wants to know how many times per hour a certain strand of virus reproduces. Suppose a sample of 23 viruses is drawn and observed that the mean is 7.4. The sample yields the standard deviation of 1.8. Construct a 95% confidence interval for the mean number of reproductions per hour.

> n=23

> MEAN=7.4

> CT=qt(0.975,22)

> ST=1.8

> round(c(MEAN-CT\*ST/sqrt(n), MEAN+CT\*ST/sqrt(n)),2)

[1] 6.62 8.18

4) The data frame barley in in lattice package and contains the yield, variety, year and site, giving the barley yields(bushels/acre) in 1931 and 1932 for 10 varieties of barley grown at six sites.

a) Construct Normal Q-Q plot of the barley yield

> library(lattice)

> attach(barley)

> qqnorm(barley$yield)

> qqline(barley$yield)

b) Construct a 95% confidence interval for mu the mean barley yield in 1932.

> x=subset(barley,year==1932)

> mu<-mean(x$yield)

> CT<-qt(.975,22)

> ST<-sd(x$yield)

> round(c(mu-CT\*ST/sqrt(23), mu+CT\*ST/sqrt(23)),2)

[1] 27.71 35.82

# 6.Hypothesis Testing

**When normal distribution：**

> t.test(x$yield,alternative="greater",mu=40)

> t.test(Caliper1,Caliper2, alt="less",conf.level=0.05,paired=T，var.equal=TRUE)

**For Varance and Proportion(Lecture 14)**

Test the hypothesis that the variance is greater to 10.

> library(TeachingDemos)

> x=c(27 , 26 ,31 , 32 , 30 , 28, 26 , 24 , 31 , 30 ,23 , 30 , 23)

> sigma.test(x,sigmasq=10，alternative="greater",conf.level=0.9)#(For one sample sd)

(b) Test for the equality of the variances at.

> var.test(x1,x2)#(For two samples sd)

> binom.test(50,500,alternative="greater",p=0.05)#for pro

> prop.test(50,500, alt="greater", p=0.05, correct=TRUE)

# 7.Chi-square Test(Lecture 15)

**Goodness of Fit**

**Independence**

H0 : the two variables are independent

Ha : the two variables are dependent

**Homogeneity**

# 8.Test for Normality

The Kolmogorov-Smirnov and Shapiro-Wilks tests for

normality calculate the probability that the sample was drawn from a

normal population.

The hypotheses used are:

*H*0 : The sample data are not significantly different than a normal population

*Ha* : The sample data are significantly different than a normal population

It will be a good idea to perform the exploratory data analysis using

*EDA()* function available in the PASWR package.

> ks.test(x, "pnorm", mean=5, sd=2)

>shapiro.test(x)

>lillie.test function is located in nortest package

# 9.Not normal:

Test to see whether the median waiting time is less than 6 minutes.

>x<-c(8,2.1,3.8,8.6,7.3,6.1,1.4,2.9,5.5,2.7,4.8,4.6,1,8.7,0.8)

> wilcox.test(x,mu=6, alternative="less")

*H*0: The two sample come from same distribution

I *Ha*: The two samples come from different distributions

> A=c(5.8, 1.0, 1.1, 2.1, 2.5, 1.1, 1.0, 1.2, 3.2, 2.7)

> B=c(1.5, 2.7, 6.6, 4.6, 1.1, 1.2, 5.7, 3.2, 1.2, 1.3)

> wilcox.test(A,B)

1. Do an appropriate test if there is a difference in the mean height.

> wilcox.test(galton$child,galton$parent)

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

Decision: Reject the null hypothesis concluding that there is a difference

between times for the first trial and time for the second trial.

# 10.Correlation

1. Calculate the Pearson correlation coefficient.

>cor.test(data$Score,data$Hours)

1. Calculate the Spearman correlation Coefficient

> cor.test(data$Score,data$Hours,method="spearman")

1. Calculate the Kendall’ s Tau.

> cor.test(data$Score,data$Hours,method="kendall")