STAT 5309 – SPRING 2022

LAB 0

***CONTENTS: R Basics –Statistics: Population distribution- Sampling distribution(Normal, ChiSquare, t-distributions, F-distribution)**
***Due: Thurs, Jan 27**

A. PRACTICE

Practice the commands in Rstudio

----R as a calculator

```
>log(2)          #natural log
>log(2, base=10)
>exp(0.6)
```

##----Reading data into R

#------Reading small data files: c() and scan()

```
>data1 <- c(1,2,3, 4,5,6) ; data1       # a vector of numerics
>data2 <- c("Mon", "Tues", "Wed", "Thurs", "Fri", "Sat", "Sun")    #a vector of characters

>data3 <- scan()               #enter data as numeric, commas are not needed
>data4 <- scan(what = "character")               #enter text
>scan(sep =",")                          #data are numeric, separated by  commas,
>scan(sep="," , what ="char")   #data are character , separated by commas.
```

#---Read from a directory

```
>setwd("C:/Users/USER/Documents/Data ")    #make default directory;   use forward
                                                                slashes.
>getwd()      #default directory
```

#------Read Excel files: read.csv(), read.delim()   (saved as .csv (comma separated version))

```
> data<- read.csv(file.choose(), sep=","   ,  header=TRUE)
> data <- read.csv(file.choose(), sep=","   ,  header=TRUE,  row.names =1)
```

------Read text file: read.table() (file save as .txt)

```
>data<- read.table(file.choose(), header=TRUE)
>data
```

Note: Missing data are denoted as NA's (not available)

-------Objects vectors, matrices , data frames of numeric data, character data;

 # ------Vectors (column vectors); Matrices

```
>numbers <- c(1,2,3,4,5)
>class(numbers)
```

```
[1] "numeric"                                # vector of numerics
 > days <- c("Mon", "Tues", "Wed")
> days
[1] "Mon"  "Tues" "Wed"                # vector of characters,  in " "
> class(days)
```

#------Convert numeric data to character data ;

```
>numbers_char  <- as.character(numbers)
>numbers_char
[1] "1" "2" "3" "4" "5"
```

#-------Matrices: A matrix is a two-dimensional data object, *with rows and columns.* A matrix can be a single row or a single column.

```
> row1<- c(1,2,3,4)
> row2  <- c(3,4,5,6)
> row1
[1] 1 2 3 4                        #a column vector
```

```
#Form a matrix using vectors:  cbind() , rbind()
> mat1 <-  cbind(row1, row2)             #cbind():  combine vectors as columns
> mat1
     row1 row2
[1,]    1    3
[2,]    2    4
[3,]    3    5
[4,]    4    6



> mat2 <- rbind(row1, row2)             #rbind(): combine vectors as rows
> mat2
     [,1] [,2] [,3] [,4]
row1    1    2    3    4
row2    3    4    5    6

> mult <- mat2  %*%    mat1       #  multiply 2 matrics
> mult
```

#--------Data frames: a data frame is a two-dimensional object,(which looks like the matrix object). All data frames are rectangular. R handles data file in table form.

(To perform matrix operations (addition/subtraction/multiplication/transpose/inverse), we require the data frames converted into matrix object)

```
> class(airfares)
[1] "data.frame"                     #airfares is a data frame
>airfares
  City Fare Distance
1    1  360     1463
2    2  360     1448
3    3  207      681
```

```
4    4  111      270
5    5   93      190


> dim(airfares)                          #dimension of the data frame
[1] 17  3
> colnames(airfares)
[1] "City"     "Fare"     "Distance"
> rownames(airfares)
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14" "15" "16" "17"

 > airfares$Fare                          #retrieve the 2nd column of data frame
 [1] 360 360 207 111  93 141 291 183 309 300  90 162 477  84 231  54 429
```

----View objects in R: all items you stored in the R environment are called **objects**.

```
>ls()                    #retrieve all items stored;  all items should be saved with a name
>ls(pattern="Ad")    #retrieve all which contain "Ad"
>rm(grades)            #remove  "grades"
```

#--- attach(), detach()-------------

```
> attach(airfares)                         # attach the data frame to the R environment
> Fare                                     # $ symbol  is not needed
 [1] 360 360 207 111  93 141 291 183 309 300  90 162 477  84 231  54 429


>detach(airfares)                       #detach the data frame from the R environment
>airfares_mat   <-  as.matrix(airfares)    #convert data frame airfares as a matrix
>airfares <- data.frame(airfares_mat)     #convert back to data frame
```

#-----------Combine vectors into data frames:

```
> days<- c("Mon", "Tues", "Wed", "Thurs")     #days  is a vector of characters
> sales <- c(2,3,4,1)
> mysales <- data.frame(rbind(days, sales))
> mysales

          X1   X2    X3    X4
days    Mon Tues Wed Thurs
sales    2      3      4     1

   # R assigns column (variable) names: X1, X2, X3, X4

> mysales_2 <- data.frame(cbind(days, sales))
> mysales_2
```

```
      days sales
1      Mon        2
2     Tues        3
3      Wed        4
4     Thursday    1
```

R not assign column names , since column names already exist

#---------------Factors

```
>height <- c(132,151,162,139,166,147,122)
>weight <- c(48,49,66,53,67,52,40)
>gender <- c("male","male","female","female","male","female","male")
>data <- data.frame( cbind(height, weight, gender))
>gender.fact<- factor(gender)
```

```
  height weight gender
1    132     48   male
2    151     49   male
3    162     66 female
4    139     53 female
5    166     67   male
6    147     52 female
7    122     40   male
```

B. **PROBABILITY DISTRIBUTIONS:** PDF (Prob. Density Function); CDF(Cumulative Distribution Function).

CDF is the area under the PDF ,from left to x.

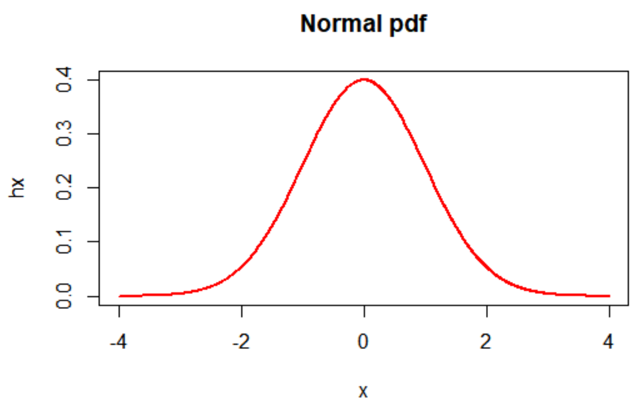
1. **Normal distribution**

Normal PDF

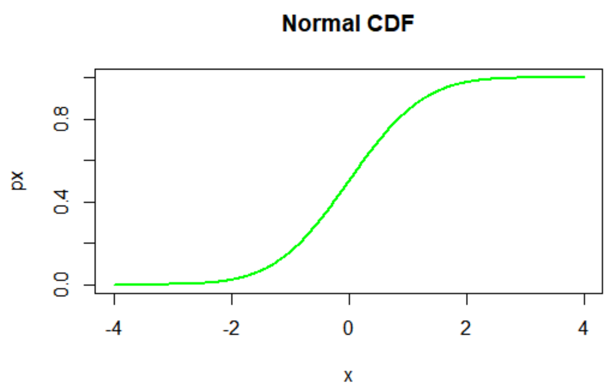
$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\ e\wedge\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Normal PDF and CDF: use dnorm(); pnorm()

```
x <- seq(-4, 4, .01)
hx <-dnorm(x, 0,1)
plot(x, hx, type="l", lwd=2,  col="red",  main="Normal pdf")
```



#normal cdf
x <- seq(-4, 4, .01)
px <-pnorm(x, mean=0,sd=1)
plot(x, px, type="l", lwd=2, col="green", main="Normal CDF")



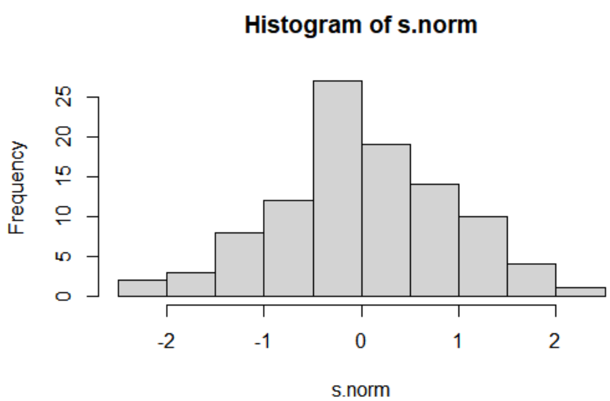
# OR use function dnorm()
curve(dnorm(x), from=-4, to=+4, main="Normal pdf")
curve(pnorm(x), from=-4, to=4)

To check normal population using random sample: (a) Sample histogram (b) Sample boxplot (c) QQ Plot (d) Shapiro test.

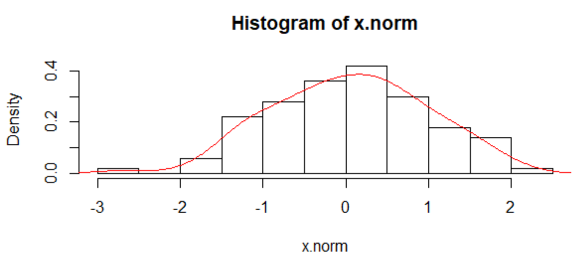
(a) Histogram: a **random** sample from normal population will likely has bell-shaped histogram;

```
s.norm <- rnorm(100)
hist(s.norm)                    #histogram on frequency

hist(s.norm, probability=TRUE)    # histogram on probability
```

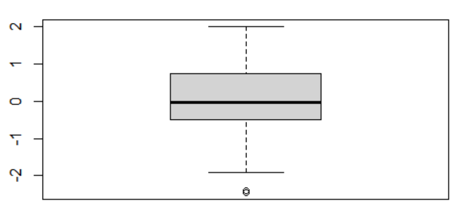



```
points(density(s.norm), col="red", type="l")   # add points(line) using empirical density
```



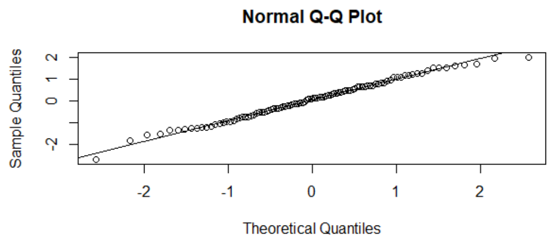
(b) Box plot

```
boxplot(s.norm)
```



(c) Normal Probability Plot

qqnorm(x.norm)
qqline(x.norm)



Note: Points lie on a straight line. Normal.

(d) Shapiro test

```
> shapiro.test(s.norm)

        Shapiro-Wilk normality test

data:  s.norm
W = 0.98773, p-value = 0.4886
```

Note: Support the Null: Normal
Population is Normal if sample is random.

Chi-Squares ; t; F are called sampling distributions(distributions coming from sampling variables)

2. Chi-squares distribution

If $X_1, X_2, \ldots, X_m$ are m independent random variables having the standard normal distribution, N(0,1) then

$$V = X_1^2 + X_2^2 + \cdots + X_m^2 \sim \chi^2_{(m)}$$

follows a Chi-Squared distribution with m degrees of freedom.

Notation: $V \sim \mathrm{X}_m^2$

Theorem:

(i)

$df = m$

$\mu = m,$

$\sigma^2 = 2m$

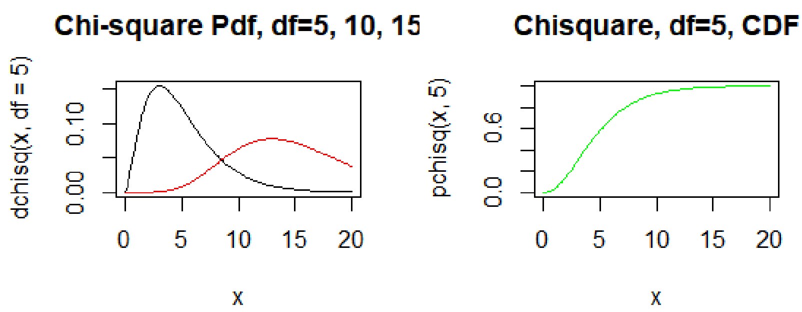
(ii) Sample Sum Squares SXX,: $SXX = (n-1)s^2$;

$\frac{SXX}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2\,(n-1)$ (the ratio between SXX and σ^2)

PDF and CDF

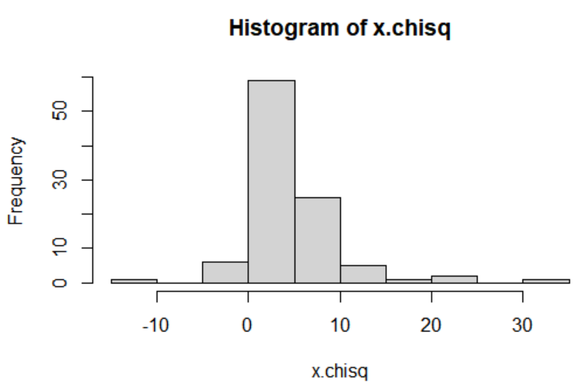
curve(dchisq(x, df=5), from=0, to=20, main="Chi-square Pdf, df=5, 10, 15")
curve(dchisq(x, 15), from=0, to=20, col="red", add=TRUE)

curve(pchisq(x, 5), from=0, to=20, col="green", main="Chisquare, df=5, CDF")

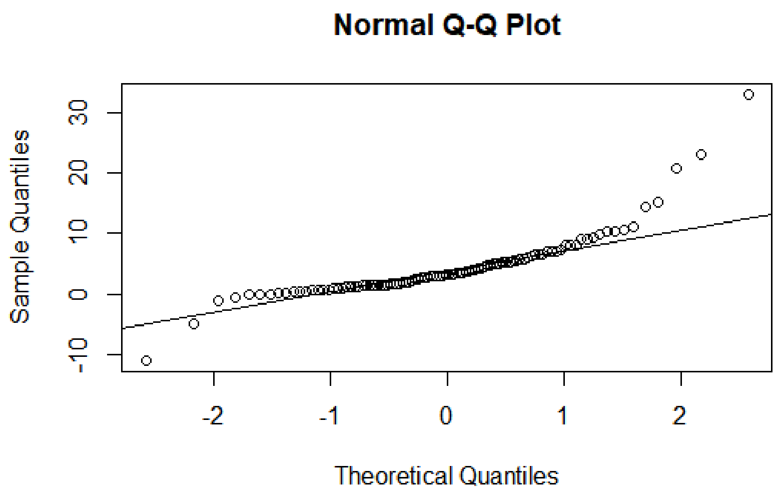


Chi-square variable: create a sample Chisquare variable, degree freedom 5

```
> abline(a=10.6035, b= 0.5949)
> abline(a=10.6035, b= 0.5949, lwd=3, col="red")
> z1 <- rnorm(100)
> z2 <- rnorm(100)
> z3 <- rnorm(100)
> z4 <- rnorm(100)
> z5 <- rnorm(100)
> x.chisq <- z1^1 +z2^2 +z3^3+z4^2+z5^2
> hist(x.chisq)
```



qqnorm(x.chisq)
qqline(x.chisq)




```
> shapiro.test(x.chisq)

        Shapiro-Wilk normality test

data:  x.chisq
W = 0.91646, p-value = 9.173e-06
```

Note: Null is Normality. Population is not Normal

3. T-distribution:

$T = \frac{Z}{\sqrt{\frac{\chi^2(n)}{n}}} \sim t_n$, t distribution, n degrees of freedom

$Z \sim N(0,1)$;
$X \sim \mathrm{X}_n^2$

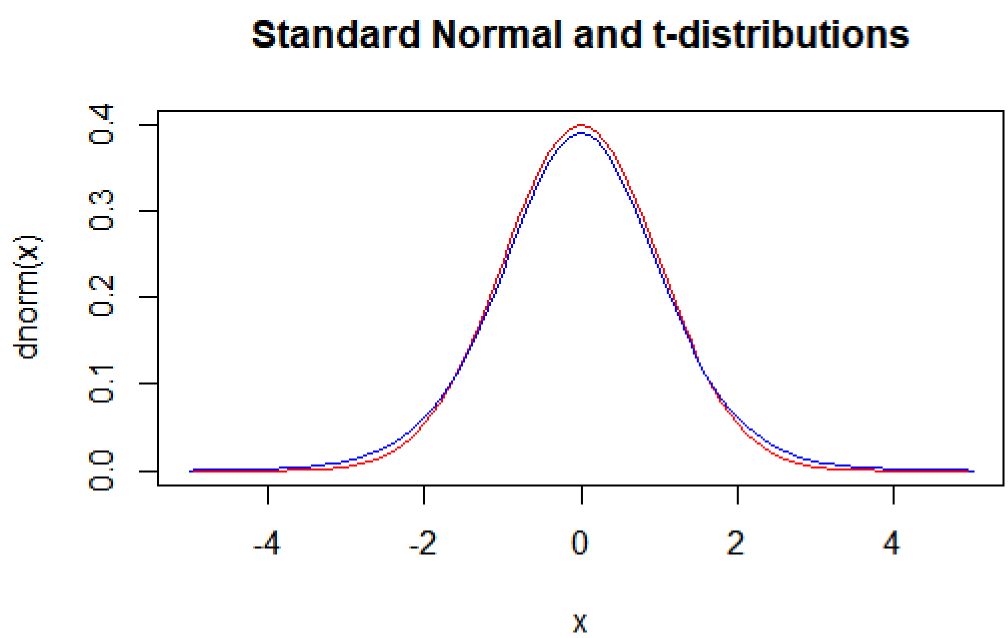
PDF:

$$f(x; n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{(n\,\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}$$

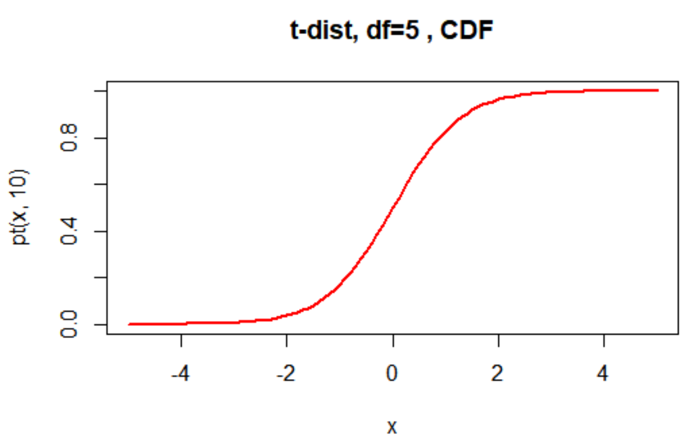
$Mean\ \ \mu = 0, \quad Variance\ \ \sigma^2 = \frac{n}{n-2}\ \ , n > 2, \quad df = n$

Plot Normal pdf and t-pdf on same plot: use dt(); pt(); qt()

curve(dnorm(x), from=-5, to=5, col="red", main="Standard Normal PDF")
curve(dt(x, 10), from=-5, to=5, col="blue", add=TRUE)

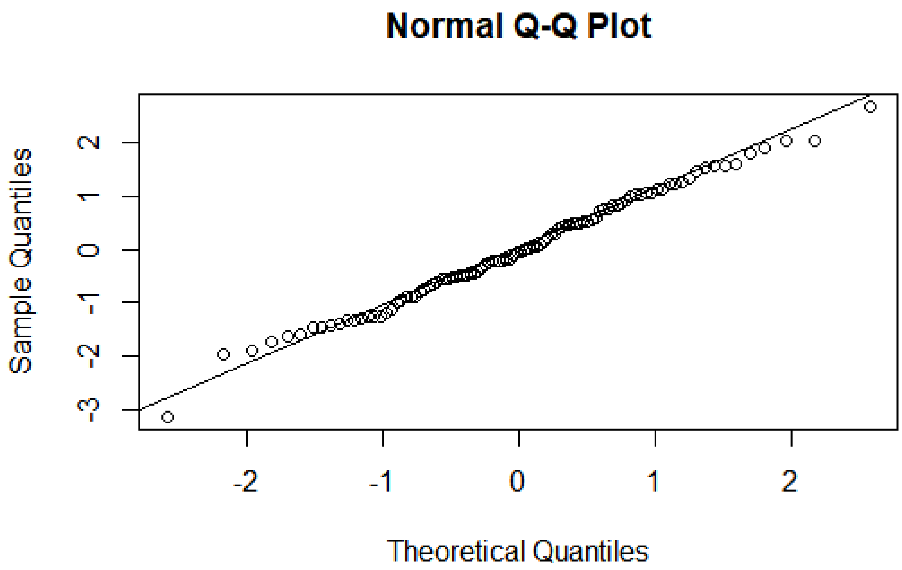


curve(pt(x, 10), from=-5, to=5,lwd=2, col="red" ,main="t-dist, df=5 , CDF")




```
## create a sample of t distribution :   rt()
x.t <- rt(100, df=10)

qqnorm(x.t)
qqline(x.t)
```



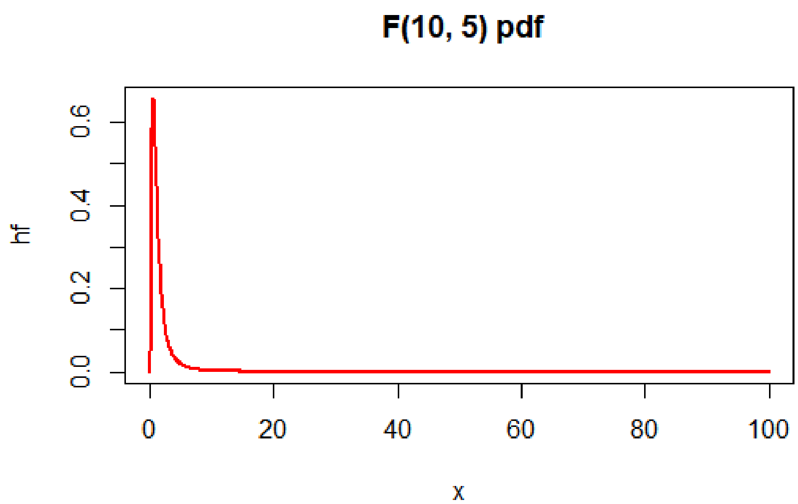
Note: Some points are off the line

4. F-distribution

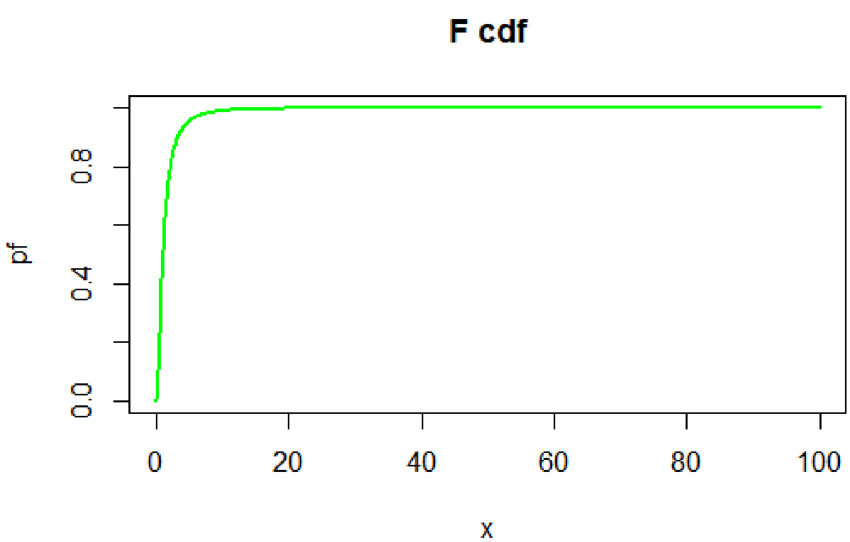
Definition: Ratio of 2 average Chi-square variables , degree freedom m,n is a F variable, Degree freedom (m,n)

$$F = \frac{\frac{X_m^2}{m}}{\frac{X_n^2}{n}} \sim F_{m,n}$$

```
x<- seq(0, 100,by=.1)
hf<- df(x, df1=10, df2=5)
plot(x, hf, type="l", lwd=2, col="red", main=" F(10, 5) pdf")
```




```
pf <- pf(x,df1=10,df2=5)
plot(x, pf, type="l",  lwd=2, col="green", main="F cdf")
```



C. Important Theorems:

1. Central Limit Theorem:

(i) If population $\boldsymbol{X} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ **,** then for all sample mean size n(fixed) $\overline{\boldsymbol{X}} \sim \boldsymbol{N}(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}})$

(ii) If population $\boldsymbol{X} \sim \boldsymbol{DISTR}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ **,** then for all sample mean size **n> 30** $\overline{\boldsymbol{X}} \sim \boldsymbol{N}(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}})$

2. Population variance $\boldsymbol{\sigma^2}$

Sample Sum Squares SXX,: $SXX = (n-1)s^2$;

$\frac{SXX}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2\ (n-1)$ (the ratio between SXX and σ^2)

D. PROBLEMS

1.

(a) Generate a random sample of 100 from t-distribtion, degree of freedom 10. Check qqnorm(); qqline(), Shapiro test. Remarks,

(b) Generate a random sample of 100 from a Chi-square distribution, df =5. Perform same procedures as in (a) . Remarks.

2.

```
sample <-c(26.4,23.5,25.4,22.9,25.2,39.2,25.5,31.9,26.0,44.6,35.5,38.6,
      30.1,31.0,30.8,32.8,47.7,39.1,55.3,50.7,73.8,71.1,68.4,77.1,
      19.4,19.3,18.7,19.0,23.2,21.3,23.2,19.9,18.9,19.8,19.6,21.9)
```

(a) Write a 95%-CI for the population mean. What assumption about population for the work, suppose the sample is random.

(b) Write a 95%- CI for population standard deviation.

3. Quantile-Quantile (QQ) Plot

Run the following code

```
sample  <-c(26.4,23.5,25.4,22.9,25.2,39.2,25.5,31.9,26.0,44.6,35.5,38.6,
      30.1,31.0,30.8,32.8,47.7,39.1,55.3,50.7,73.8,71.1,68.4,77.1,
      19.4,19.3,18.7,19.0,23.2,21.3,23.2,19.9,18.9,19.8,19.6,21.9)
hist(sample)
sample.s <-sort(sample)           #sort data increasing
rank <- rank(sample.s)            #rank data from 1 to 36
size <- length(sample.s)          # size of data

p <- (rank-.5)/size               #cummulative prob of data
z.quantile <- qnorm(p)            # Standard Normal quantiles with such probability
plot(x=z.quantile, y=sample.s, pch=16, main=”QQ Plot”)   #scatterplot of x=Z quantiles, y= data sorted

abline(lm(sample.s ~ z.quantile))
```