Probability and Statistics (UCS410)
Experiment 1: Basics of R programming

We hope you already have R and RStudio installed in your computers. Good luck with your programming in R! Let us start with the very basics of R programming here in our first Lab :)

If this is the first time, you are using R, then there is a lot to explore. You can try making codes for the following problems. This will give you a general idea of how to use R and over the next few labs you will learn more.

---

(1) Create a vector $c = [5, 10, 15, 20, 25, 30]$ and write a program which returns the maximum and minimum of this vector. vector=c() max(v) min(v)

(2) Write a program in R to find factorial of a number by taking input from user. Please print error message if the input number is negative.factorial() for(i in 1:num) fact*=i

(3) Write a program to write first n terms of a Fibonacci sequence. You may take n as an input from the user. fno=0 sno=1 for(i in 1:n) if(i==1)print fno else if(i==2)print sno else next=fno+sno fno=sno and sno=next

(4) Write an R program to make a simple calculator which can add, subtract, multiply and divide. read operator as.character

(5) Explore plot, pie, barplot etc. (the plotting options) which are built-in functions in R.

plot(vec1,vec2,main="Name of plot", xlab="Label of X", ylab="Label of Y", col="")
default: scatter
type="l" line plot

barplot bar graph
boxplot
pie
histogram (only 1 vector as input)

To take input from user:
as.integer()
as.character()

readline(prompt="Enter input: ")

If using variables inside print then add paste: print(paste("var is: ",varName))

for(i in 1:10)

varName=switch(operator,"+"=cat("sum is",a+b),""=_ _ _)

```
res=function(){
  return (___)
}
now call res()
```

Probability and Statistics (UCS410)
Experiment 2: Descriptive statistics, Sample space, definition of probability

---

vector=c(rep("Gold",20),_ _) sample(vector,no_of_random_coins_drawn,replace=T/F)

(1) (a) Suppose there is a chest of coins with 20 gold, 30 silver and 50 bronze coins. You randomly draw 10 coins from this chest. Write an R code which will give us the sample space for this experiment. (use of **sample**(): an in-built function in R)
(b) In a surgical procedure, the chances of success and failure are 90% and 10% respectively. Generate a sample space for the next 10 surgical procedures performed. (use of **prob**(): an in-built function in R) given probabilites & asked to find sample space
sample(vector,no_of_points,replace=T,prob=c(_,_))

pbirthday: computes the probability of a coincidence

qbirthday: computes the smallest number of observations needed to have a specific probability

(2) A room has n people, and each has an equal chance of being born on any of the 365 days of the year. (For simplicity, we'll ignore leap years). What is the probability that two people in the room have the same birthday? pbirthday(n for n values,365 no of days,2 ppl have same bdy)
(a) Use an R simulation to estimate this for various n.
(b) Find the smallest value of n for which the probability of a match is greater than .5. qbirthday(0.5 probab given,365 total no of days,2 two ppl bday on same date)

(3) Write an R function for computing conditional probability. Call this function to do the following problem: P(cloudy)=0.04 P(rain)=0.2 P(c|r)=0.85 so find P(c~r) P(r|c)=P(r~c)/P(c) suppose the probability of the weather being cloudy is 40%. Also suppose the probability of rain on a given day is 20% and that the probability of clouds on a rainy day is 85%. If it's cloudy outside on a given day, what is the probability that it will rain that day?

(4) The iris dataset is a built-in dataset in R that contains measurements on 4 different attributes (in centimeters) for 150 flowers from 3 different species. Load this dataset and do the following: d=iris to access specific colm: $

(a) Print first few rows of this dataset. head(d,no_of_rows_to_print) similarly use tail
(b) Find the structure of this dataset. str(d)
(c) Find the range of the data regarding the sepal length of flowers. range(d$Sepal.Length)
(d) Find the mean of the sepal length. mean(d$Sepal.Length)
(e) Find the median of the sepal length. median(d$Sepal.Length)
(f) Find the first and the third quartiles and hence the interquartile range. quantile(d$Sepal,0.25/0.75)
(g) Find the standard deviation and variance. sd(d$sepal) var(d$sepal) IQR(d$sepal)
(h) Try doing the above exercises for sepal.width, petal.length and petal.width.
(i) Use the built-in function summary on the dataset Iris. summary(d)

(5) R does not have a standard in-built function to calculate mode. So we create a user function to calculate mode of a data set in R. This function takes the vector as input and gives the mode value as output.

```
uni=unique(vec)
uni[which.max(tabulate(match(vec,uni)))]    first maps vector to uni & gets index then tabulates its occurence then finds max out of it using which.max
```

when success of each trial is known & prob of getting success in some n trials asked then use pbinom
when mean & sd r given then use pnorm
when average no of events:lambda is given & probab of observing some k intervals in finite time is asked: ppois()
when successes & failures given (defective/non-defective) & prob aksed WITHOUT REPLACEMENT: dhyper
for prob density/mass fnc...calculate prob of exact k success given success on each trial: dbinom()

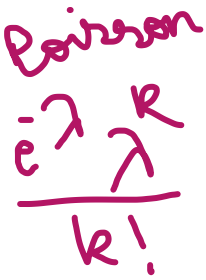
to find commulative distribution fnc: cumsum()

Probability and Statistics (UCS410)
Experiment 3: Probability distributions

---

(1) Roll 12 dice simultaneously, and let X denotes the number of 6's that appear. Calculate the probability of getting 7, 8 or 9, 6's using R. (Try using the function **pbinom**;
pbinom(no of success, total_no_of_trials, prob of suceess in each trial, lower.tail=default True ie P(x<=q) if false then P(X>q)
n=12 res=9-6 to get entries 9,8,7 so x=pbinom(9,n,1/6) y=pbinom(6,n,1/6) res=x-y
diff(pbinom(c(6,9),n,1/6))

(2) Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?
pnorm(total mrks 100,mean=72,std=15.2)-pnorm(more than 82,72,15.2,lower.tail=FALSE)



(3) On the average, five cars arrive at a particular car wash every hour. Let X count the number of cars that arrive from 10AM to 11AM, then $X \sim \text{Poisson}(\lambda = 5)$. What is probability that no car arrives during this time. Next, suppose the car wash above is in operation from 8AM to 6PM, and we let Y be the number of customers that appear in this period. Since this period covers a total of 10 hours, we get that $Y \sim \text{Poisson}(\lambda = 5 \times 10 = 50)$. What is the probability that there are between 48 and 50 customers, inclusive? ppois(no_of_events:0 ie no car arrives,lambda:5 on an average)
diff(ppois(c(47,50 , lambda:50))

(4) Suppose in a certain shipment of 250 Pentium processors there are 17 defective processors. A quality control consultant randomly collects 5 processors for inspection to determine whether or not they are defective. Let X denote the number of defectives in the sample. Find the probability of exactly 3 defectives in the sample, that is, find $P(X = 3)$. dyhper(success_in_SAMPLE: 3 defective , success_in_POPUlation: total no of defectives: 17 , failures_in_POPUL: non-defective:250-17=233 , size: collects 5 process)

(5) A recent national study showed that approximately 44.7% of college students have used Wikipedia as a source in at least one of their term papers. Let X equal the number of students in a random sample of size n = 31 who have used Wikipedia as a source. X ranges from 0 stud to 31 stud so X=0:31

(a) How is X distributed? dbinom(no_of_success:X range, no_trials: 31 , probab_of_success: 44.7%= 0.447)

(b) Sketch the probability mass function. plot this dbinom for prob mass type="h" for histogram

(c) Sketch the cumulative distribution function. cumsum(y:dbinom ans)

(d) Find mean, variance and standard deviation of X. weighted.mean(x,dbinom ans y)
var=weighted.mean((x-mean)^2,dbimon:y))
sd=sqrt(var)

---

pbinom: commulative distribution fnc
dbinom: probab density/MASS fnc

Average: sum(numbers*probabilities vector) , weighted.mean(x,p) , c(x%*%p)
To find EXPECTED VALUE of cts RANDOM VARIABLE T: integrate(fnc f, lower limit 0,upper limit:Inf)
To find EXPECTED VALUE GIVEN PROBABILITY MASS FNC DISTRIBUTION: 1.form eqn 2. sapply 3. sum()
First Moment=Mean=integrate(fnc,lower limit, upper limit)
Second Moment= first x^2 in the fnc then integrate(NEW fnc,lower limit, upper limit)
variance=second moment - (first moment)^2
sapply():apply a function to each element of a vector

Probability and Statistics (UCS410)

Experiment 4

(Mathematical Expectation, Moments and Functions of Random Variables)

1. The probability distribution of X, the number of imperfections per 10 meters of a synthetic fabric in continuous rolls of uniform width, is given as

x	0	1	2	3	4
$p(x)$	0.41	0.37	0.16	0.05	0.01

create vectors for both p & x

Find the average number of imperfections per 10 meters of this fabric.

(Try functions **sum( )**, **weighted.mean( )**, c(a **%*%** b) to find expected value/mean.
sum(x*p) weighted.mean(x,p) c(x%*%p)

2. The time T, in days, required for the completion of a contracted project is a random variable with probability density function $f(t) = 0.1\ e^{(-0.1t)}$ for $t > 0$ and 0 otherwise. Find the expected value of T. create this fnc: with()*t integrate(func,0,Inf) and to get value use $value
Use function **integrate( )** to find the expected value of continuous random variable T.

3. A bookstore purchases three copies of a book at $6.00 each and sells them for $12.00 each. Unsold copies are returned for $2.00 each. Let X = {number of copies sold} and Y = {net revenue}. If the probability mass function of X is

x	0	1	2	3
$p(x)$	0.1	0.2	0.2	0.5

create these vectors for x & p

first find revenue using eqn in a fnc: f(x): rev=12*x-6*3+2*(3-x)
y=sapply(x,rev) and then sum(y*pprobab p)

Find the expected value of Y.

4. Find the first and second moments about the origin of the random variable X with probability density function $f(x) = 0.5e^{-|x|}$, $1 < x < 10$ and 0 otherwise. Further use the results to find Mean and Variance.
fnc that stores this density fnc: x*0.5*exp(-abs(x)))
(kth moment = $E(X^k)$, Mean = first moment and Variance = second moment – Mean^2
variance=secondMoment-(firstMoment)^2
first moment=MEAN= integrate(fnc,ll:1,up:10) print its $value
second moment=first create another fnc that has x^2 instead of x...integrate(new fnc,1,10) print $value

5. Let X be a geometric random variable with probability distribution

$$f(x) = \frac{3}{4}\left(\frac{1}{4}\right)^{x-1}, x = 1,2,3,\ldots$$ define fnc as fnc(x): 3/4*(1/4)^(x-1)

Write a function to find the probability distribution of the random variable $Y = X^2$ and find probability of Y for X = 3. Further, use it to find the expected value and variance of Y for X = 1,2,3,4,5.

prob dist of Y when x=3: fnc(3)

then find expected value=sum(given x^2* fnc(x)) == mean

second_moment=sum(x^4& fnc(x))
var=second_moment- (exp value)^2

punif
dexp
pexp
rexp
dgamma
pgamma
qgamma

Probability and Statistics (UCS410)

Experiment 5

(Continuous Probability Distributions)

1. Consider that X is the time (in minutes) that a person has to wait in order to take a flight. If each flight takes off each hour X ~ U(0, 60). Find the probability that

   (a) waiting time is more than 45 minutes, and 1-punif(45,min=0,max=60)

   (b) waiting time lies between 20 and 30 minutes. punif(30,min=0,max=60)-punif(20,min=0,max=60)

2. The time (in hours) required to repair a machine is an exponential distributed random variable with parameter $\lambda = 1/2$.

   (a) Find the value of density function at x = 3. dexp(3,rate=1/2)

   (b) Plot the graph of exponential probability distribution for $0 \le x \le 5$. Generate a sequence seq(0,5,by=0.1) dexp(sq,1/2) plot(seq,dexp)

   (c) Find the probability that a repair time takes at most 3 hours. pexp(3,rate=1/2)

   (d) Plot the graph of cumulative exponential probabilities for $0 \le x \le 5$. Generate a sequence seq(0,5,by=0.1) pexp(sq,1/2) plot(seq,pexp)

   (e) Simulate 1000 exponential distributed random numbers with $\lambda = ½$ and plot the simulated data. rexp(1000,rate=1/2) hist(rexp)

3. The lifetime of certain equipment is described by a random variable X that follows Gamma distribution with parameters $\alpha = 2$ and $\beta = 1/3$. new_rate=1/beta

   (a) Find the probability that the lifetime of equipment is (i) 3 units of time, and (ii) at least 1 unit of time. pgamma(1,shape=alpha,rate=new_rate) dgamma(3,shape=alpha,rate=new_rate)

   (b) What is the value of c, if $P(X \le c) \ge 0.70$? (**Hint:** try quantile function qgamma()) qgamma(0.70,shape=alpha,rate=new_rate)

School of Mathematics(SOM)

Thapar Institute of Engineering and Technology (TIET)

Probability and Statistics(UCS410) Exp. sheet 06 (Joint probability mass and density functions)

```
install.packages("pracma")
library("pracma")
```

(1) The joint probability density of two random variables X and Y is

```
fnc1=function(x,y){     (2*(2*x+3*y))/5   }
```

$$f(x,y) = \begin{cases} 2(2x+3y)/5; & 0 \le x, y \le 1 \\ 0; & elsewhere \end{cases}$$

Then write a R-code to

```
ans1=integral2(  fnc1,  xmin=0,  xmax=1,  ymin=0,  ymax=1  )
ans1$Q
```

(i) check that it is a joint density function or not? (Use integral2())

```
fnc2=function(y){   fnc1(1,y)   }
gx=integral(fnc2,0,1)
```

(ii) find marginal distribution $g(x)$ at $x = 1$.

(iii) find the marginal distribution $h(y)$ at $y = 0$.

```
fnc4=function(x,y){   x*y*fnc1(x,y)   }
ans=integral2(fnc4,   xmin=0,  xmax=1,  ymin=0,  ymax=1)
```

(iv) find the expected value of $g(x,y) = xy$.

(2) The joint probability mass function of two random variables X and Y is

matrix

$$f(x,y) = \{(x+y)/30;\ \ x = 0,1,2,3;\ \ y = 0,1,2\}$$

```
x=c(0:3)    y=c(0:2)
```

Then write a R-code to

```
matrix(  c(  function1(0,y),function1(1,y),function1(2,y),function1(3,y)  ),  nrow=4,  ncol=3,  byrow=T  )
```

(i) display the joint mass function in rectangular (matrix) form.

```
sum(M)     if(sum(M)==1){  print("Yes
```

(ii) check that it is joint mass function or not? (use: Sum())

```
apply(M matrix,1,sum)
```

(iii) find the marginal distribution $g(x)$ for $x = 0,1,2,3$. (Use:apply())

```
apply(M matrix,2,sum)
```

(iv) find the marginal distribution $h(y)$ for $y = 0,1,2$. (Use:apply())

```
matrix M[1,2] / marginal of y [2]
```

(v) find the conditional probability at $x = 0$ given $y = 1$.

(vi) find $E(x), E(y), E(xy), Var(x), Var(y), Cov(x,y)$ and its correlation coefficient.

```
e_x1=sum(x* marginal of x)

e_y1=sum(y* marginal of y)

e_x2=sum(x*x* marginal of x)
e_y2=sum(y*y* marginal of y)

var_x=e_x2- (e_x1*e_x1)
var_y=e_y2- (e_y1*e_y1)


e_xy=0
for (i in 1:length(x)) {
  for (j in 1:length(y)) {
    e_xy=e_xy + x[i] * y[j] * M[i, j]
  }
}
cov_xy= e_xy-e_x1*e_y1

corr_coef= cov_xy/(sqrt(var_x*var_y))
```

School of Mathematics(SOM)

Thapar Institute of Engineering and Technology (TIET)

Probability and Statistics(UCS410) Exp. sheet 07 (Chi-square, t-distribution, F-distribution)

---

(1) Use the $rt(n, df)$ function in r to investigate the t-distribution for $n = 100$ and $df = n - 1$ and plot the histogram for the same.

```
t_values=rt(n, df)
hist(t_values,breaks = 10)
```

(2) Use the $rchisq(n, df)$ function in r to investigate the chi-square distribution with $n = 100$ and $df = 2, 10, 25$.

```
chi_sq_df2 = rchisq(n, df = 2)
chi_sq_df10 = rchisq(n, df = 10)
chi_sq_df25 = rchisq(n, df = 25)
hist( chi_sq_df_ )
```

(3) Generate a vector of 100 values between -6 and 6. Use the $dt()$ function in r to find the values of a t-distribution given a random variable x and degrees of freedom 1,4,10,30. Using these values plot the density function for students t-distribution with degrees of freedom 30. Also shows a comparison of probability density functions having different degrees of freedom (1,4,10,30).

(4) Write a r-code

```
x=seq(  -6, 6, length.out = 100   )
plot(x, density_df30, type = "l")
density_df1=dt(x sequence  , df = 1)
lines(x, density_df1, col = "red", lwd = 2)
legend("topright", legend = c("df=1", "df=4", "df=10", "df=30"), col = c("red", "green", "purple", "blue"),lwd=2,cex=0.5)
```

(i) To find the 95^{th} percentile of the F-distribution with $(10, 20)$ degrees of freedom.

(ii) To calculate the area under the curve for the interval $[0, 1.5]$ and the interval $[1.5, +\infty)$ of a F-curve with $v_1 = 10$ and $v_2 = 20$ (USE $pf()$).

(iii) To calculate the quantile for a given area (= probability) under the curve for a F-curve with $v_1 = 10$ and $v_2 = 20$ that corresponds to $q = 0.25, 0.5, 0.75$ and 0.999. (use the $qf()$)

(iv) To generate 1000 random values from the F-distribution with $v_1 = 10$ and $v_2 = 20$ (use $rf()$)and plot a histogram.

```
p=0.95
res=qf(p,  10,  20)

area_0to1.5=  pf(1.5,10,20)
area1.5_toinf=   1-pf(1.5,10,20)

quant0.25=qf(0.25,10,20)
quant0.75=qf(0.75,10,20)

randomvals=rf(1000,10,20)
hist(randomvals,breaks=10)
```

Probability and Statistics (UCS410)
Experiment 8

- A pipe manufacturing organization produces different kinds of pipes. We are given the monthly data of the wall thickness of certain types of pipes (data is available on LMS Clt-data.csv).

  The organization has an analysis to perform and one of the basic assumption of that analysis is that the data should be normally distributed.

  You have the following tasks to do:

  (a) Import the csv data file in R.

  (b) Validate data for correctness by counting number of rows and viewing the top ten rows of the dataset.

  (c) Calculate the population mean and plot the observations by making a histogram.

  (d) Mark the mean computed in last step by using the function abline.

```
data=read.csv(file.choose())
head(data,10)
dim(data)
population_mean=mean(data$Wall.Thickness)
hist(data$Wall.Thickness)
abline(v=population_mean)
```

  See the red vertical line in the histogram? That's the population mean. Comment on whether the data is normally distributed or not?

  Now perform the following tasks:

  (a) Draw sufficient samples of size 10, calculate their means, and plot them in R by making histogram. Do you get a normal distribution.

  (b) Now repeat the same with sample size 50, 500 and 9000. Can you comment on what you observe.

  Here, we get a good bell-shaped curve and the sampling distribution approaches normal distribution as the sample sizes increase. Therefore, we can recommend the organization to use sampling distributions of mean for further analysis.

```
draw_sample_means = function(sample_size) {
 s=c()
 n=9000
 for(i in 1:n){
  s[i]=mean(sample(population_mean  ,  sample_size,  replace=TRUE))
 }
 # Plot the histogram of sample means
 hist(s,main = paste("Histogram of Sample Means (Sample Size =", sample_size, ")"),
     xlab = "Sample Mean",
     ylab = "Frequency")
 abline(v=mean(s),col="yellow")
}

draw_sample_means(10)
```

Q2: Plot the scatter diagram and a regression line that will enable us to predict Cholesterol level on age. Further, estimate the cholesterol level of a 60 year-old man.

```
plot(age, cholesterol)
regressionLine=lm ( cholesterol~age  )
abline(  regressionLine, col = "red", lwd = 2)
predicted_cholesterol=predict (regressionLine, data.frame(age = 60)  )
```

Q3. Assume that the differences between the pre-course and post-course test scores are normally distributed, and a high score on the test indicates a strong level of assertiveness. Do the collected data, at 5% level of significance, provide enough evidence to conclude that research scholars become more assertive after completing the course

```
t_test_result=t.test  (after, before, paired = TRUE, alternative = "greater")
t  test  result
```