STK 210: Practical 5

1 PROC IML: Probability Distributions and Density Functions

1.1 The PDF, CDF and QUANTILE function

1.1.1 PDF function

Returns a value from a probability distribution or a probability density.

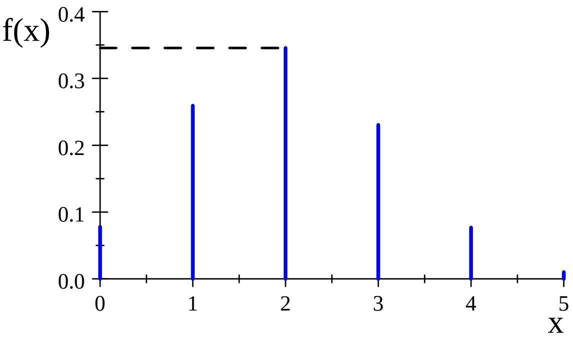
Syntax:

```
PDF(distribution,quantile<,parameter-1,...,parameter-k>)
```

Example

1.

$$X \sim BIN(5, 0.4)$$



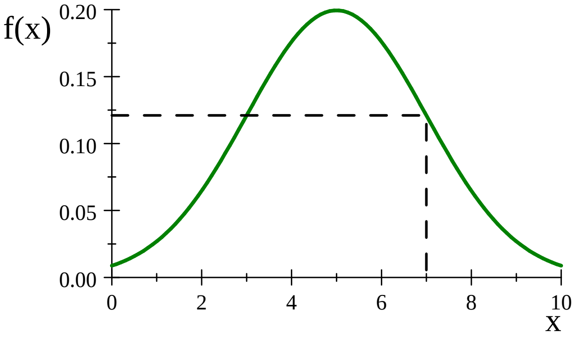
Note: $P(X = 2) = 0.345$

```
pdf('Binomial',2,0.4,5)
```

```
0.3456
```

2.

$$X \sim N(5, 4)$$



Note: $P(X = 7) = 0$ Continuous variable!

```
pdf('Normal',7,5,2)
```

```
0.1209854
```

1.1.2 CDF function

Returns a value from a cumulative probability distribution.

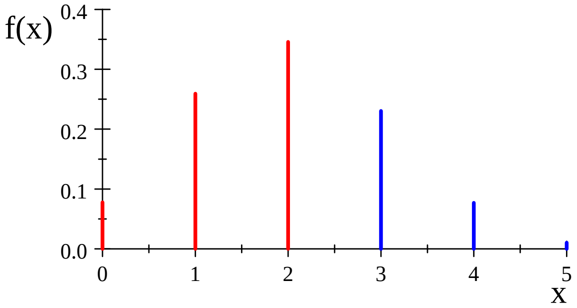
Syntax:

```
CDF(distribution, quantile <, parameter-1, ..., parameter-k>)
```

Example

1.

$$X \sim BIN(5, 0.4)$$



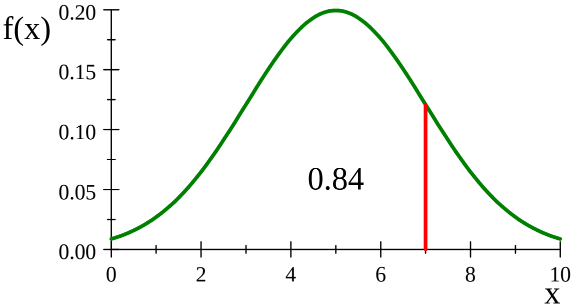
Note: $P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.68256$

```
cdf('Binomial',2,0.4,5)
```

```
0.68256
```

2.

$$X \sim N(5, 4)$$



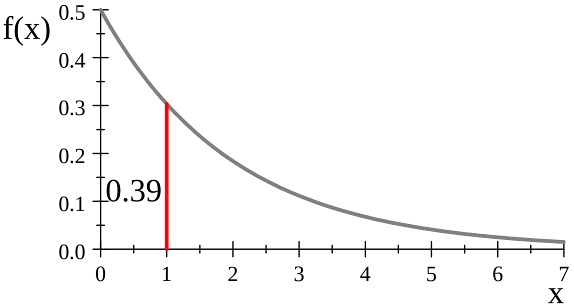
Note : $P(X \le 7) = 0.8413447$

```
cdf('Normal',7,5,2)
```

```
0.8413447
```

3.

$$X \sim EXP(2)$$



Note : $P(X \le 1) = 0.3934693$

```
cdf('Exponential',1,2)
```

```
0.3934693
```

1.1.3 QUANTILE function

Returns the quantile from a distribution when you specify the left probability (CDF).

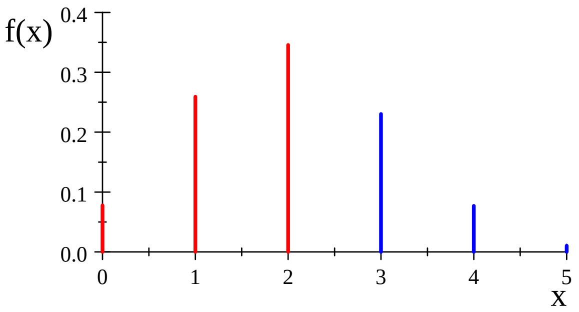
Syntax:

```
QUANTILE(distribution,probability,parameter-1,...,parameter-k)
```

Example

1.

$$X \sim BIN\,(5, 0.4)$$



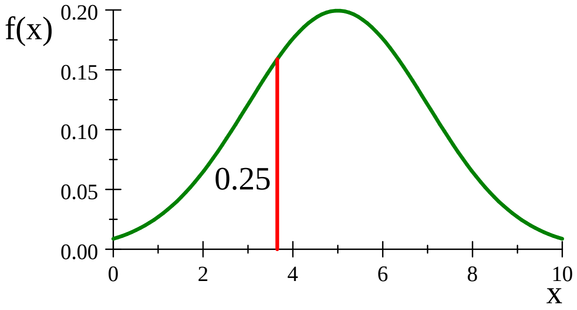
Note: $me = Q_2 = 2$

```
quantile('Binomial',0.5,0.4,5)
```

```
2
```

2.

$$X \sim N\,(5, 4)$$



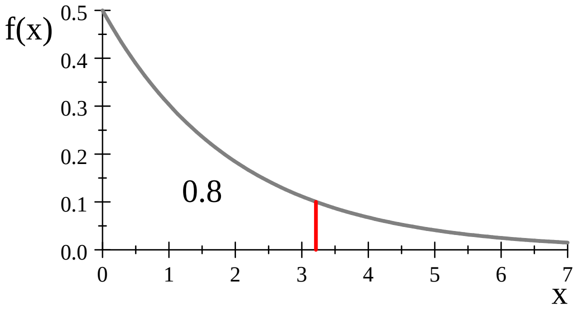
Note : $\Phi^{-1}\,(0.25) = 3.651 = Q_1$

```
quantile('Normal',0.25,5,2)
```

```
3.6510205
```

3.

$$X \sim EXP\,(2)$$



Note : $P_{80} = 3.2188758$

```
quantile('Exponential',0.8,2)
```

```
3.2188758
```

SAS Program:

```
proc iml;
pdf1=pdf('Binomial',2,0.4,5);
pdf2=pdf('Normal',7,5,2);
cdf1=cdf('Binomial',2,0.4,5);
cdf2=cdf('Normal',7,5,2);
cdf3=cdf('Exponential',1,2);
quant1=quantile('Binomial',0.5,0.4,5);
quant2=quantile('Normal',0.25,5,2);
quant3=quantile('Exponential',0.8,2);
print pdf1 pdf2, cdf1 cdf2 cdf3, quant1 quant2 quant3;
```

SAS Output:

```
     pdf1      pdf2
   0.3456 0.1209854

     cdf1      cdf2      cdf3
  0.68256 0.8413447 0.3934693

   quant1    quant2    quant3
        2 3.6510205 3.2188758
```

1.2 The RANDFUN function

The RANDFUN function returns a matrix of random numbers from a specified distribution.

Syntax:

```
RANDFUN (N, 'Distribution' <, param1> <, param2> <, param3>);
```

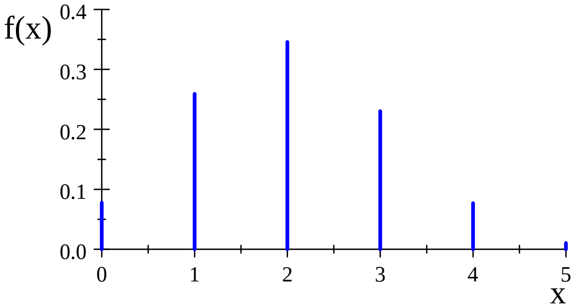
If N is a positive integer, the function returns an $(N \times 1)$ column vector of random numbers that are drawn from the Distribution family with the specified parameters. If N is a vector that contains a pair of integers, the function returns an $(N[1] \times N[2])$ matrix of random numbers.

1.2.1 Example

The following example simulates data from six distributions:

- The binomial distribution with probability $p = 0.4$ and $n = 5$.

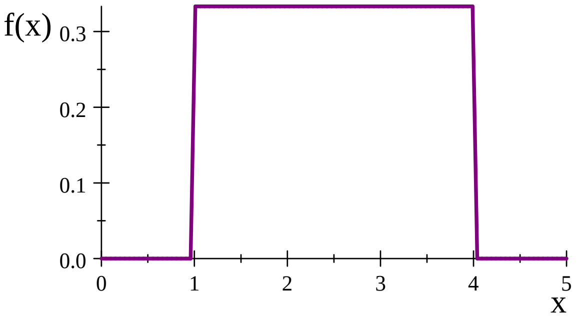
$$X \sim BIN(5, 0.4)$$



```
randfun({4,2},'Binomial',0.4,5)
```

- The uniform distribution on the interval $(1, 4)$.

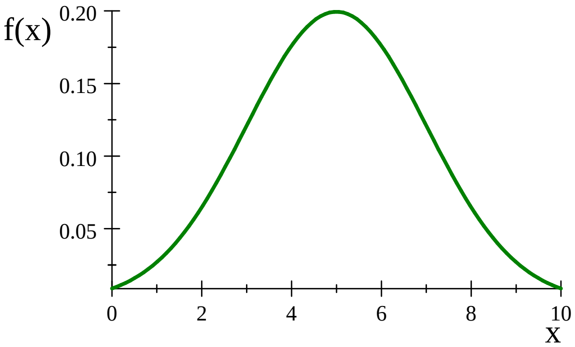
$$X \sim UNIF(1, 4)$$



```
randfun(5,'Uniform',1,4)
```

- The normal distribution with mean 5 and standard deviation 2.

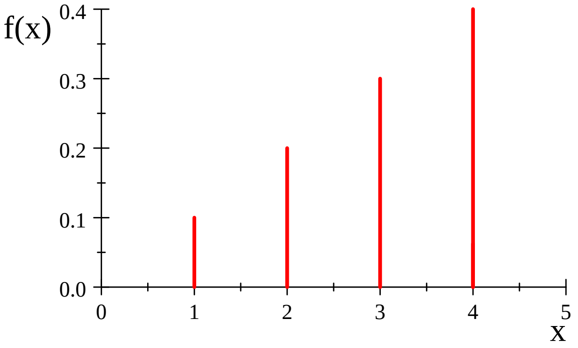
$$X \sim N(5, 4)$$



```
randfun({6,3},'Normal',5,2)
```

- Probability distribution with pdf

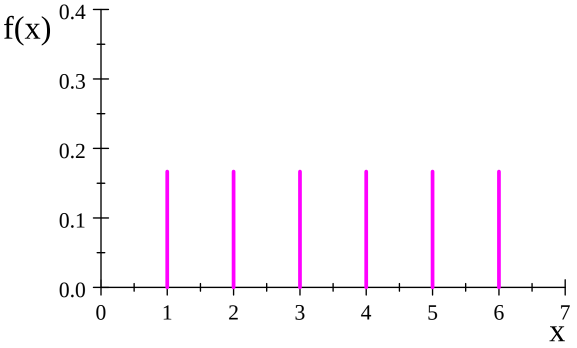
x	$f(x)$
1	0.1
2	0.2
3	0.3
4	0.4



```
randfun(8,'Table',{0.10.20.30.4})
```

- Discrete unifirm distribution with $k = 6$. (Six sided die)

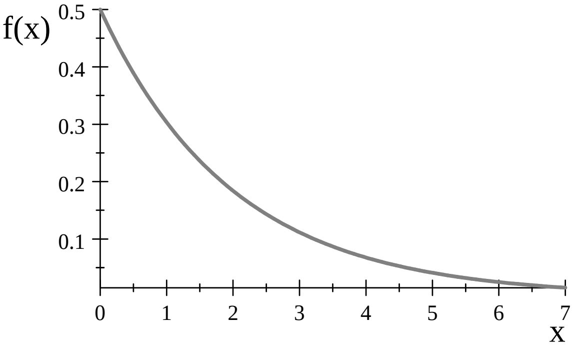
$$X \sim DU(6)$$



```
randfun(10,'Table',J(1,6,1)/6)
```

- The exponential distribution with expected value $\theta = 2$.

$$X \sim EXP(2)$$



```
randfun(10,'Exp',2)
```

SAS Program:

```
proc iml;
call randseed(111,1); b=randfun({4 2},'Binomial',0.4,5);
call randseed(222,1); u=randfun(5,'Uniform',1,4);
call randseed(333,1); x=randfun({6,3},'Normal',5,2);
call randseed(444,1); t=randfun(8,'Table',{0.1 0.2 0.3 0.4});
call randseed(555,1); d=randfun(10,'Table',J(1,6,1)/6);
call randseed(666,1); e=randfun(10,'Exponential',2);
y=round(x,0.01);
print 'Binomial' b 'Uniform' u;
print 'Normal' x y;
print  'pdf' t 'Discrete Uniform' d 'Exponential' e e[format=5.3];
quit;
```

Note: The domain of the random variable of the `Table` distribution is always $1, 2, 3, 4, \ldots$ i.e. starting from one and increasing by one. Any other values for the domain will require a transformation of the independent variable.

SAS Output:

```
                 b                           u
Binomial         1         2 Uniform 2.4609738
                 5         1         3.6192201
                 2         3         1.0589588
                 2         1          1.734596
                                     2.7172301

               x                             y
Normal 7.1163233 7.0181356 5.9939514      7.12      7.02      5.99
        7.161919 8.3331989 10.421176      7.16      8.33     10.42
       5.6373503 6.1226537 6.8433468      5.64      6.12      6.84
        4.986196 5.9922651 3.0798055      4.99      5.99      3.08
       4.6550445 2.0904793 6.6324223      4.66      2.09      6.63
       5.8148733 5.8685608 5.9153982      5.81      5.87      5.92

            t                          d                     e     e
pdf         2 Discrete Uniform         4 Exponential 0.9211326 0.921
            4                          5             1.9367753 1.937
            4                          2             0.6606601 0.661
            4                          4             9.9203163 9.920
            1                          1             3.5870836 3.587
            4                          5             1.1537839 1.154
            2                          6             2.2273609 2.227
            4                          2              3.385624 3.386
                                       5             0.2886844 0.289
                                       1             1.4764443 1.476
```

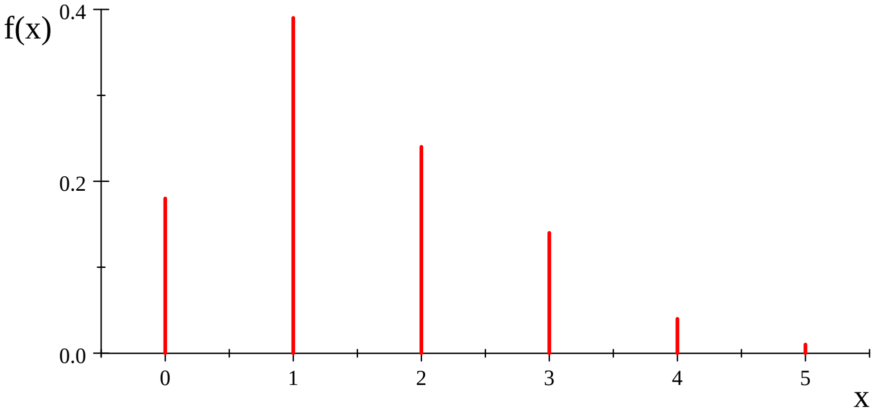
2 Exercise: Probability Distributions

1. Let

$$X = \text{ number of automobiles sold during a day at DiCarlo Motors}$$

The probability distribution is

x	0	1	2	3	4	5
$f(x)$	0.18	0.39	0.24	0.14	0.04	0.01



(a) Create the following two vectors in PROC IML

$$\mathbf{x} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix} \quad \text{and} \quad \mathbf{p} = \begin{pmatrix} 0.18 \\ 0.39 \\ 0.24 \\ 0.14 \\ 0.04 \\ 0.01 \end{pmatrix}$$

Print $\mathbf{x}$ and $\mathbf{p}$.

Note:

- You can use the statement `xt=0:5;` to create the numerical row vector $\mathbf{x}'$. The numerical column vector $\mathbf{x}$ can be obtained by taking the transponent `x=xt‘;`
- You can use the statement `char_x=’0’:’5’;` to create the character vector `char_x` to label the levels of the vector $\mathbf{x}$.

(b) Calculate the following probabilities by making use of submatrices.

i. $P(X = 3)$ 0.14

ii. $P(X < 4)$ 0.95

iii. $P(2 \leq X < 5)$ 0.42

iv. $P(X \geq 3)$ 0.19

Note: You may use the SUM function where necessary.

(c) Use the `CUSUM` function in PROC IML to create the vector

$$\mathbf{F} = \begin{pmatrix} F(0) \\ F(1) \\ F(2) \\ F(3) \\ F(4) \\ F(5) \end{pmatrix} = \begin{pmatrix} P(X \leq 0) \\ P(X \leq 1) \\ P(X \leq 2) \\ P(X \leq 3) \\ P(X \leq 4) \\ P(X \leq 5) \end{pmatrix} = \begin{pmatrix} 0.18 \\ 0.57 \\ 0.81 \\ 0.95 \\ 0.99 \\ 1 \end{pmatrix}$$

where $F(x)$ is the distribution function of X. Print the vector $\mathbf{F}$.

- The `CUSUM` function returns a matrix with the cumulative sums obtained by summing the elements of the matrix.
- **Syntax:**

```
CUSUM(Matrix)
```

(d) Use the vector with cumulative probabilities, $\mathbf{F}$, and calculate the value of the following:

i. me 1

ii. Q_3 2

iii. P_{95} 3

iv. $P(X = 3)$ 0.14

v. $P(X < 4)$ 0.95

vi. $P(2 \leq X < 5)$ 0.42

vii. $P(X \geq 3)$ 0.19

(e) Use the RANDSEED CALL with a seed of $\boxed{101}$ to generate a sample of size $n = 8$ from the population. Print the generated sample of size 8.

Remember: When using the RANDFUN function the domain of the `TABLE` distribution starts at one and not zero. It is therefore necessary to subtract 1 from the generated sample by making use of

```
Q1e=v-J(8,1,1);
```

(f) Use the RANDSEED CALL with a seed of $\boxed{242}$ to generate a sample of size $n = 1000$ from the population. Create the SAS dataset $\boxed{\text{DiCarlo}}$ with the variable SOLD with the 1000 observations.

i. Print the 500th observation of the generated sample of size 1000. 0

ii. Use PROC UNIVARIATE to calculate the following values from the empirical distribution of X.

A. $\min$ 0

B. $\max$ 5

C. P_{10} 0

D. Q_1 1

E. me 1

F.	Q_3	2
G.	P_{90}	3
H.	IQR (Interquartile Range)	1
I.	Range	5

iii. Use the HISTOGRAM statement to plot the empirical distribution of X by making use of the following statement in PROC UNIVARIATE

```
histogram / midpoints=0,1,2,3,4,5 interbar=10 cfill=red;
```

Note: You have to include the statement

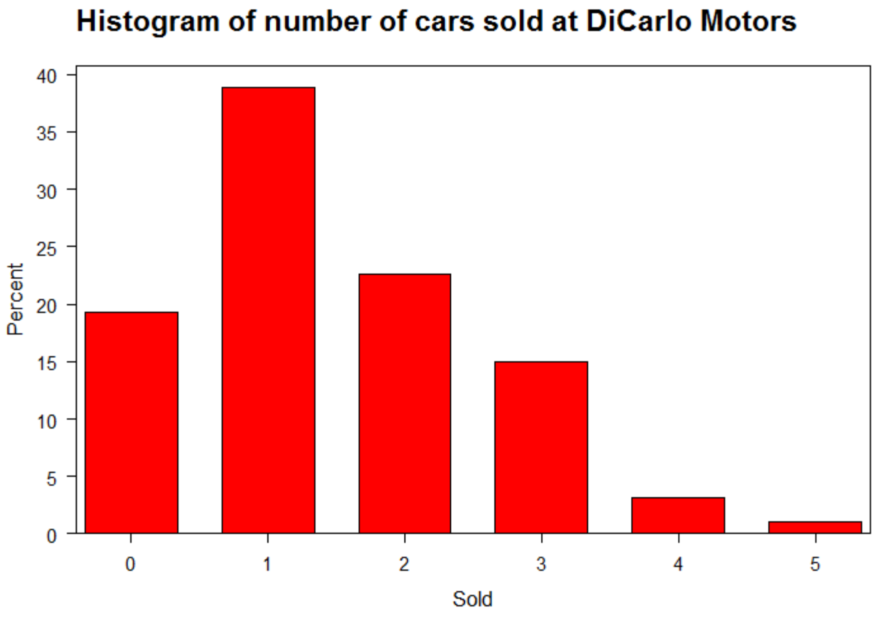
```
ods graphics off;
```

before you invoke PROC UNIVARIATE with the HISTOGRAM statement. See SAS Program below.

SAS Program:

```
ods graphics off;
proc univariate data=DiCarlo;
var sold;
histogram / midpoints=0,1,2,3,4,5 interbar=5 cfill=red;
title ’Histogram of number of cars sold at DiCarlo Motors’;
run;
```

Graph:



iv. Use PROC FREQ to determine the empirical distribution of X.

A. The empirical probability distribution of X is

x	0	1	2	3	4	5
$f(x)$	0.193		0.226		0.031	0.011

Complete!

B. Give the values of the following empirical cumulative probabilities

$$\mathbf{F} = \begin{pmatrix} F(0) \\ F(1) \\ F(2) \\ F(3) \\ F(4) \\ F(5) \end{pmatrix} = \begin{pmatrix} P(X \leq 0) \\ P(X \leq 1) \\ P(X \leq 2) \\ P(X \leq 3) \\ P(X \leq 4) \\ P(X \leq 5) \end{pmatrix} = \begin{pmatrix} P(X \leq 0) \\ P(X \leq 1) \\ P(X \leq 2) \\ P(X \leq 3) \\ P(X \leq 4) \\ P(X \leq 5) \end{pmatrix} = \begin{pmatrix} 0.193 \\ \\ 0.808 \\ \\ 0.989 \\ 1 \end{pmatrix}$$

C. Give the probabilities from the empirical distribution for

- $P(X = 3) = f(3)$ 0.150
- $P(X < 4) = F(3)$ 0.958

NB: Compare the empirical distributions with the theoretical distributions.

v. Calculate the empirical probabilities of

A. $P(2 \leq X < 5)$ 0.407

B. $P(X \geq 3)$ 0.192

by calculating the following two variables

```
data prob; set DiCarlo;
if 2<=sold<5 then grp1=1; else grp1=0;
if sold>=3   then grp2=1; else grp2=0;
```

Use PROC FREQ and PROC MEANS to obtain the two empirical probabilities.

NB: Compare the empirical probabilities with the theoretical probabilities in Question 1(b).

2. Consider an experiment in which we roll a pair of dice.

- Let:

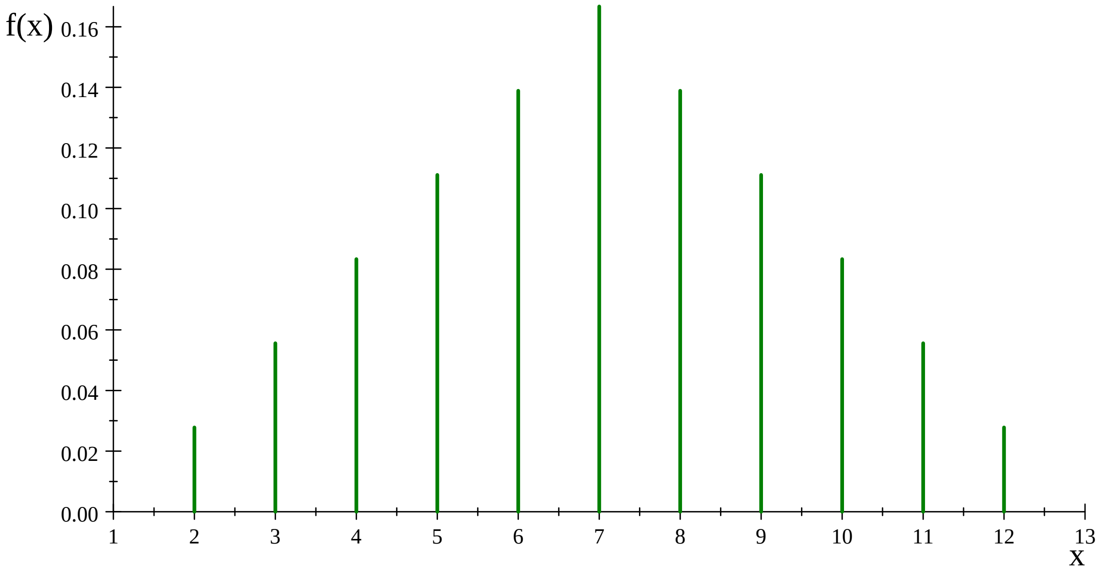
$$X = \text{Total number of points with pair of dice.}$$

- The values of X is:

Green	Red die					
die	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

- The **range** of X is $\{2, 3, 4, \ldots, 12\}$.
- The probability distribution of X is:

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
	.0278	.0556	.0833	.1111	.1389	.1667	.1389	.1111	.0833	.0556	.0278



(a) The theoretical probability distribution can be obtained from the mathematical model

$$f(x) = \frac{6 - |x - 7|}{36} \quad \text{for } x = 2, 3, \ldots, 12 \tag{1}$$

i. Create the numerical vector

$$\mathbf{x} = \begin{pmatrix} 2 \\ 3 \\ 4 \\ \vdots \\ 12 \end{pmatrix}$$

and use the vector $\mathbf{x}$ to create the vector with probabilities

$$f(\mathbf{x}) = \begin{pmatrix} f(2) \\ f(3) \\ f(4) \\ \vdots \\ f(12) \end{pmatrix} \tag{2}$$

by making use of (1).

ii. Give the value of the following:

A. $P(X = 7)$ 0.1667

B. $P(X = 4)$ 0.0833

C. $P(X = 10)$ 0.0833

D. mode 7

iii. Use the SUM function to calculate the following probabilities:

A. $P(5 \leq X \leq 7)$ 0.4167

B. $P(4 < X < 10)$ 0.6667

C. $P(6 \leq X < 9)$ 0.4444

D. $P(X \geq 4)$ 0.9167

E. $P(X \leq 9)$ 0.8333

iv. Use the `CUSUM` function to create the vector of cumulative probabilities

$$F(\mathbf{x}) = \begin{pmatrix} F(2) \\ F(3) \\ F(4) \\ \vdots \\ F(12) \end{pmatrix}$$

v. Use the cumulative probabilities to calculate the following

A. $P(X = 8)$ 0.13889

B. $P(X > 3)$ 0.91667

C. $P(X < 6)$ 0.27778

D. $P(3 < X < 7)$ 0.33333

E. $P(6 \leq X \leq 8)$ 0.44444

F. P_{10} 4

G. Q_1 5

H. me 7

I. Q_3 9

J. P_{95} 11

(b) Generate the matrix $\mathbf{A}$ with $n = 10$ rows (10 times that we roll pair of dice) and $t = 2$ columns (pair of dice), by making use of the statement

```
A=randfun({10,2},'TABLE',p);
```

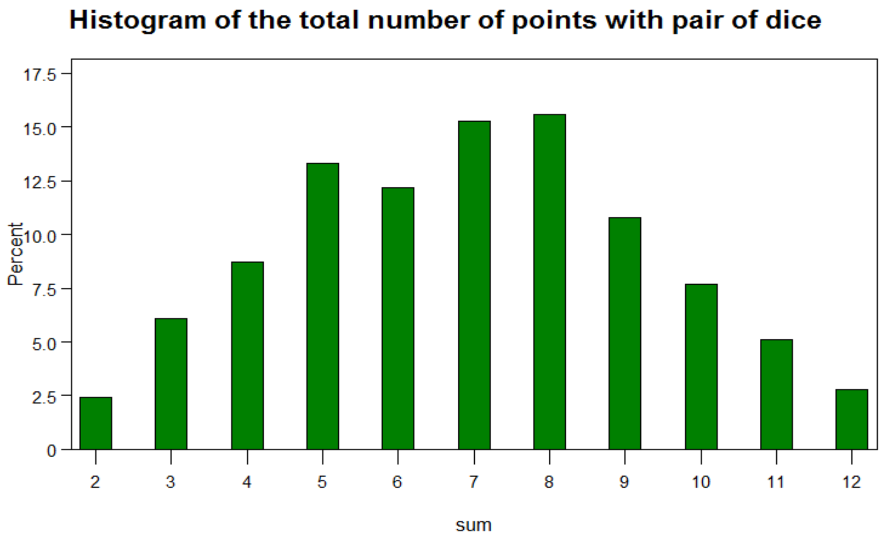
Use the RANDSEED CALL with a seed of $\boxed{615}$ to generate the sample. Create the vector $\mathbf{b}$ with the sum of the two rolls of the pair of dice. Print the matrix $\mathbf{A}$ and the vector $\mathbf{b}$.

(c) Use the RANDSEED CALL with a seed of $\boxed{823}$ to generate the sample of size $n = 1000$ from the population of X. Create the SAS data set DICE with the variable SUM denoting the sum of the two rolls of the pair of dice.

i. Use PROC UNIVARIATE to obtain the following empirical values for X:

A. $\min$	2
B. $\max$	12
C. mode	8 (differs from theoretical distribution)
D. P_{10}	4
E. Q_1	5
F. me	7
G. Q_3	9
H. P_{95}	11

ii. Use the HISTOGRAM statement in PROC UNIVARIATE to draw the empirical distribution of X.



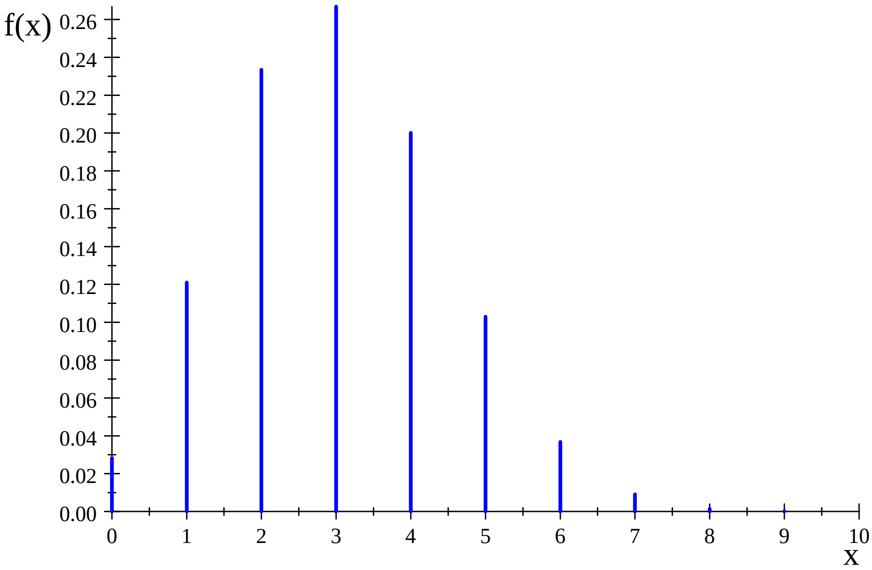
iii. Use PROC FREQ or PROC MEANS to calculate the following probabilities from the empirical distribution of X.

A. $P(X = 8)$	0.156
B. $P(X > 3)$	0.915
C. $P(X < 6)$	0.305
D. $P(3 < X < 7)$	0.342
E. $P(6 \leq X \leq 8)$	0.431

3. Suppose $n = 10$ customers enter the Martin Clothing Store. The probability that any one customer will make a purchase is $p = 0.3$. Let

$$X = \text{ number of customers making a purchase}$$

then $X \sim BIN\,(10, 0.3)$. **Graph:**



(a) Give the mode. 3

(b) Give the range. 10

(c) Calculate the mean and standard deviation of X by making use of

i. $\mu = np$ 3

ii. $\sigma = \sqrt{np\,(1-p)}$ 1.449

(d) Determine the value of the following by making use of the PDF, CDF and QUANTILE functions.

Syntax:

```
PDF('BINOMIAL',m,p,n)
CDF('BINOMIAL',m,p,n)
QUANTILE('BINOMIAL',probability,p,n)
```

where

$$\begin{aligned} n &= \text{number of independent trials} \\ p &= \text{probability of a success} \\ m &= \text{number of successes} \\ \texttt{probability} &= \text{for specific quantile} \end{aligned}$$

(See p.1-4.)

i. $P\,(X = 3)$ 0.2668

ii. $P\,(3 < X \leq 7)$ 0.3488

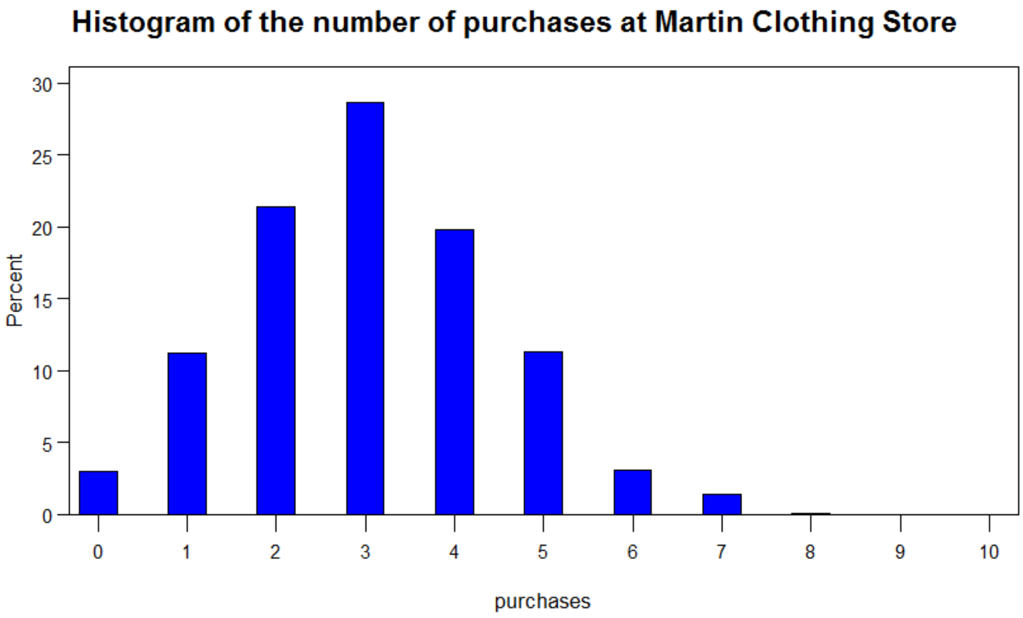
iii. $P\left(X \geq 3\right)$	0.6172
iv. $F\left(4.5\right)$	0.8497
v. P_{05}	1
vi. me	3
vii. Q_1	2
viii. P_{90}	5

(e) Use the RANDSEED CALL with a seed of $\boxed{615}$ to generate a sample of size $n = 1000$ from the population. Create the SAS data set MARTIN with the variable PURCHASES

i. Use PROC UNIVARIATE to give the empirical values for the following:

A. μ	3.05
B. σ	1.4524
C. mode	3
D. range	8
E. P_{05}	1
F. me	3
G. Q_1	2
H. P_{90}	5

ii. Use the HISTOGRAM statement in PROC UNIVARIATE to draw the empirical distribution of X.



iii. Use PROC FREQ or PROC MEANS to calculate the following probabilities from the empirical distribution of X.

A. $P\left(X = 3\right)$	0.287
B. $P\left(3 < X \leq 7\right)$	0.356
C. $P\left(X \geq 3\right)$	0.644
D. $F\left(4.5\right)$	0.841