Sardar Patel Institute of Technology,Mumbai
Department of Electronics and Telecommunication Engineering
T.E. Sem-V (2018-2019)
ETL54-Statistical Computational Laboratory

Lab-2: Probability Distributions

Name:Manish D'Silva **Roll No.15**

Objective:To compute probability density function (pdf) and cumulative distribution function (cdf)

Outcomes:

1. To list and describe the well-known probability distributions with their characteristics.
2. To compute the probability distributions which frequently occur in Statistical Study

System Requirements: Ubuntu OS with R and RStudio installed

Introduction to Probability distribution:

A probability distribution describes how the values of a random variable is distributed. There are two types of probability distributions: Discrete and Continuous
Well-known probability distributions which are frequently occurred in statistical study:

- Binomial Distribution
- Poisson Distribution
- Continuous Uniform Distribution
- Exponential Distribution
- Normal Distribution
- Chi-squared Distribution
- Student t Distribution
- F Distribution

R Functions for Probability Distributions:

Every distribution that R handles has four functions. There is a root name, for example, the root name for the normal distribution is norm. This root is prefixed by one of the letters
p for "probability", the cumulative distribution function (c. d. f.)
q for "quantile", the inverse c. d. f.
d for "density", the density function (p. f. or p. d. f.)
r for "random", a random variable having the specified distribution
For the normal distribution, these functions are pnorm, qnorm, dnorm, and rnorm. For the binomial distribution, these functions are pbinom, qbinom, dbinom, and rbinom. And so forth.
For a continuous distribution (like the normal), the most useful functions for doing

problems involving probability calculations are the "p" and "q" functions (c. d. f. and inverse c. d. f.), because the the density (p. d. f.) calculated by the "d" function can only be used to calculate probabilities via integrals and R doesn't do integrals.

For a discrete distribution (like the binomial), the "d" function calculates the density (p. f.), which in this case is a probability

f(x) = P(X = x)

and hence is useful in calculating probabilities.

R has functions to handle many probability distributions. The table below gives the names of the functions for each distribution.

Table-1:Probability Distributions

Distribution	Functions			
Binomial	pbinom	qbinom	dbinom	rbinom
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy
Chi-Square	pchisq	qchisq	dchisq	rchisq
Exponential	pexp	qexp	dexp	rexp
F	pf	qf	df	rf
Gamma	pgamma	qgamma	dgamma	rgamma
Geometric	pgeom	qgeom	dgeom	rgeom
Hypergeometric	phyper	qhyper	dhyper	rhyper
Logistic	plogis	qlogis	dlogis	rlogis
Log Normal	plnorm	qlnorm	dlnorm	rlnorm
Normal	pnorm	qnorm	dnorm	rnorm
Poisson	ppois	qpois	dpois	rpois
Student t	pt	qt	dt	rt
Uniform	punif	qunif	dunif	runif
Weibull	pweibull	qweibull	dweibull	rweibull

Procedure:

1. Open RStudio
2. Go to RConsole (>)
3. Probability distribution in R

```
>help(rnorm) #The normal Distribution
>help(dbinom) # The Binomial Distribution
```

Probability Distributions in R:

In R, probability functions take the form

[dpqr]distribution_abbreviation ()

where the first letter refers to the aspect of the distribution returned:

d = density

p = distribution function

q = quantile function

r = random generation (random deviates)

1. Binomial Distribution

The binomial distribution is a discrete probability distribution. It describes the outcome of n independent trials in an experiment. Each trial is assumed to have only two outcomes, either success or failure. If the probability of a successful trial is p, then the probability of having x successful outcomes in an experiment of n independent trials is as follows.

$$f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{(n-x)} \quad where\ x = 0, 1, 2, ..., n$$

Problem

Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

Example Solution:

Since only one out of five possible answers is correct, the probability of answering a question correctly by random is 1/5=0.2. We can find the probability of having exactly 4 correct answers by random attempts as follows.

```
> dbinom(4, size=12, prob=0.2)
[1] 0.1329
```

To find the probability of having four or less correct answers by random attempts, we apply the function dbinom with x = 0,…,4.

```
> dbinom(0, size=12, prob=0.2) +
+ dbinom(1, size=12, prob=0.2) +
+ dbinom(2, size=12, prob=0.2) +
+ dbinom(3, size=12, prob=0.2) +
+ dbinom(4, size=12, prob=0.2)
[1] 0.9274
```

Alternatively, we can use the cumulative probability function for binomial distribution pbinom.

```
> pbinom(4, size=12, prob=0.2)
[1] 0.92744
```

Answer: The probability of four or less questions answered correctly by random in a twelve question multiple choice quiz is 92.7%.

2. Poisson Distribution

The Poisson distribution is the probability distribution of independent event occurrences in an interval. If λ is the mean occurrence per interval, then the probability of having x occurrences within a given interval is:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad where\ x = 0, 1, 2, 3, \ldots$$

Problem

If there are twelve cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.

Function used : ppois()

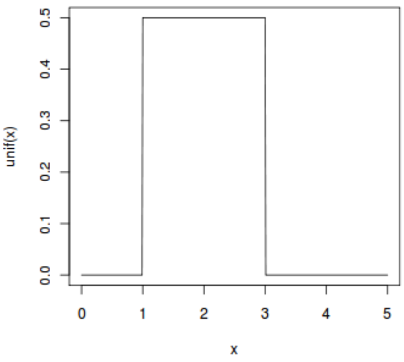
Answer: 10.1291% percent chances of having 17 or more cars crossing a bridge.

3. Continuous Uniform Distribution

The continuous uniform distribution is the probability distribution of random number selection from the continuous interval between a and b. Its density function is defined by the following.

$$f(x) = \begin{cases} \frac{1}{b-a} & when\ a \leq x \leq b \\ 0 & when\ x < a\ or\ x > b \end{cases}$$

Here is a graph of the continuous uniform distribution with a = 1, b = 3.



Problem

Select ten random numbers between one and three.

Function used : rnorm() , set.seed()(Used to get identical random numbers over all systems)

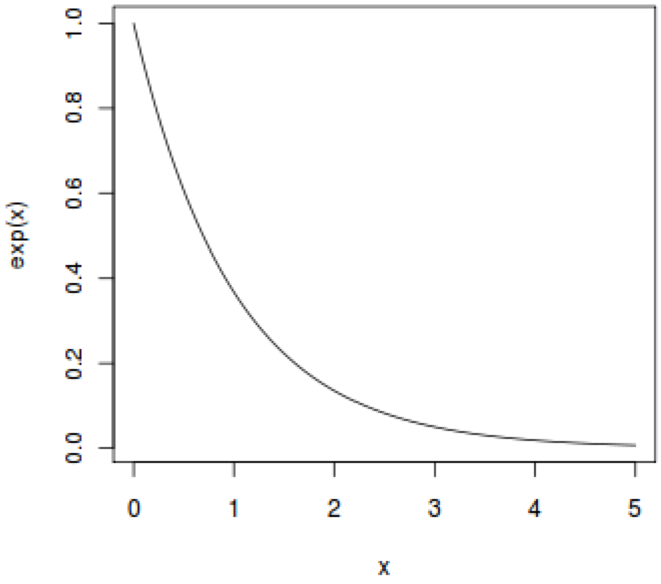
Answer:1.575155 2.576610 1.817954 2.766035 2.880935 1.091113 2.056211 2.784838 2.102870 1.913229

4. **Exponential Distribution**

The exponential distribution describes the arrival time of a randomly recurring independent event sequence. If μ is the mean waiting time for the next event recurrence, its probability density function is:

$$f(x) = \begin{cases} \frac{1}{\mu} e^{-x/\mu} & when\ x \geq 0 \\ 0 & when\ x < 0 \end{cases}$$

Here is a graph of the exponential distribution with μ = 1.



Problem

Suppose the mean checkout time of a supermarket cashier is three minutes. Find the probability of a customer checkout being completed by the cashier in less than two minutes.

Function Used : pexp()

Answer:48.65829% chances of a customer checkout being completed in less than 2mins

5.Normal Distribution

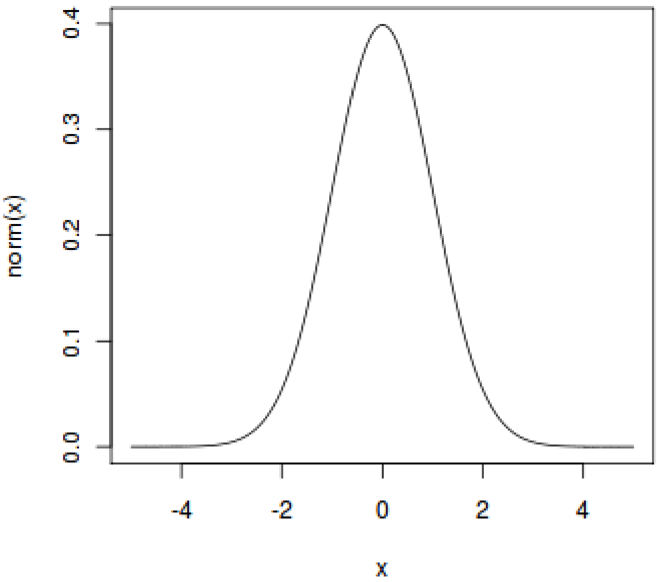
The normal distribution is defined by the following probability density function, where μ is the population mean and σ^2 is the variance.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

If a random variable X follows the normal distribution, then we write:

$$X \sim N(\mu, \sigma^2)$$

In particular, the normal distribution with μ = 0 and σ = 1 is called the standard normal distribution, and is denoted as N(0,1). It can be graphed as follows.



The normal distribution is important because of the Central Limit Theorem, which states that the population of all possible samples of size n from a population with mean μ and variance σ^2 approaches a normal distribution with mean μ and σ^2⁄n when n approaches infinity.

Problem

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?

Function Used : pnorm()

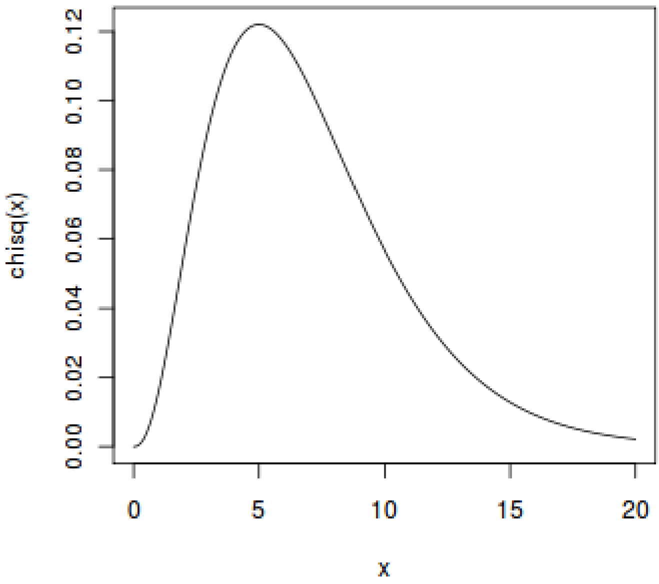
Answer:21.49176% of students score more than 84 in the exam.

6.Chi-squared Distribution

If X_1,X_2,…,X_m are m independent random variables having the standard normal distribution, then the following quantity follows a Chi-Squared distribution with m degrees of freedom. Its mean is m, and its variance is 2m.

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

Here is a graph of the Chi-Squared distribution 7 degrees of freedom.



Problem

Find the 95th percentile of the Chi-Squared distribution with 7 degrees of freedom.

Function Used: qchisq()

Answer: 14.06714 is the 95th percentile.

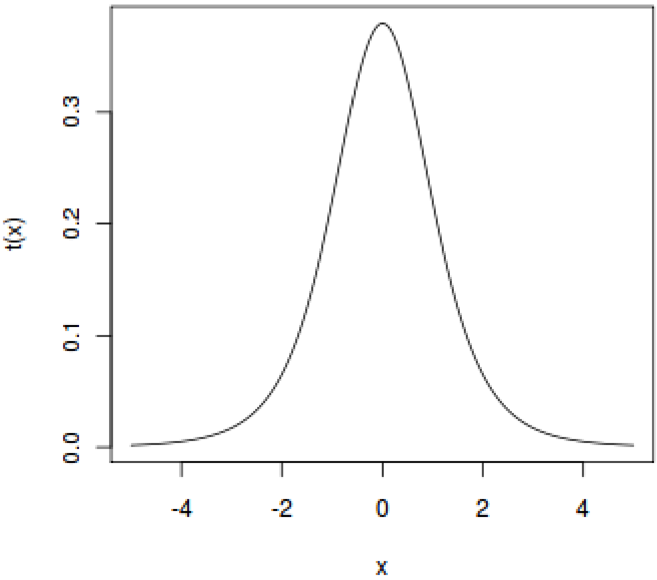
8.Student t Distribution

Assume that a random variable Z has the standard normal distribution, and another random variable V has the Chi-Squared distribution with m degrees of freedom.

Assume further that Z and V are independent, then the following quantity follows a Student t distribution with m degrees of freedom.

$$t = \frac{Z}{\sqrt{V/m}} \sim t_{(m)}$$

Here is a graph of the Student t distribution with 5 degrees of freedom.



Problem

Find the 2.5th and 97.5th percentiles of the Student t distribution with 5 degrees of freedom.

Function Used : qt()

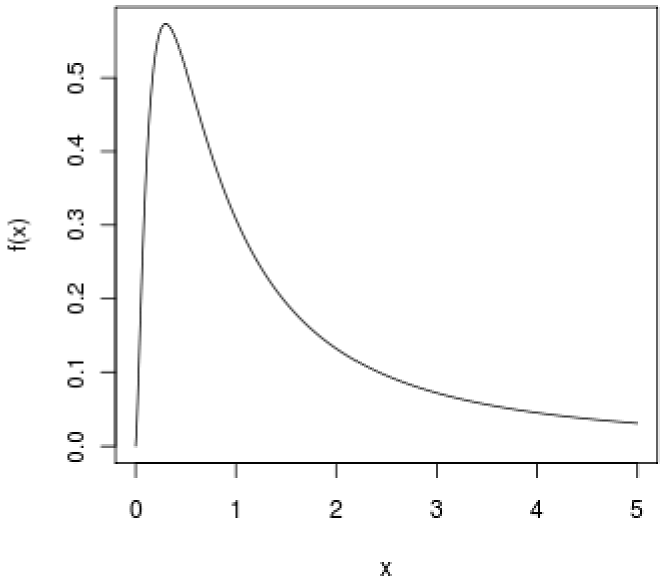
Answer:-2.570582 is the 2.5th percentile and 2.570582 is the 97.5th percentile

8. F Distribution

If V $_1$ and V $_2$ are two independent random variables having the Chi-Squared distribution with m_1 and m_2 degrees of freedom respectively, then the following quantity follows an F distribution with m_1 numerator degrees of freedom and m_2 denominator degrees of freedom, i.e., (m_1,m_2) degrees of freedom.

$$F = \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1,m_2)}$$

Here is a graph of the F distribution with (5, 2) degrees of freedom.



Problem

Find the 95th percentile of the F distribution with (5, 2) degrees of freedom.

Function Used: qf()

Answer: 19.29641 is the 95th percentile

Describe the following with respect to probability distributions:

1.

x <- rnorm(1000, mean=100, sd=15)

Here we have created a vector of 1000 numbers with mean 100 and standard deviation of 15.Thus it takes 1000 numbers with the above restrictions.

head(x) = 125.72597 106.91374 81.02408 89.69721 93.31507 118.36123

hist(x, probability=TRUE)

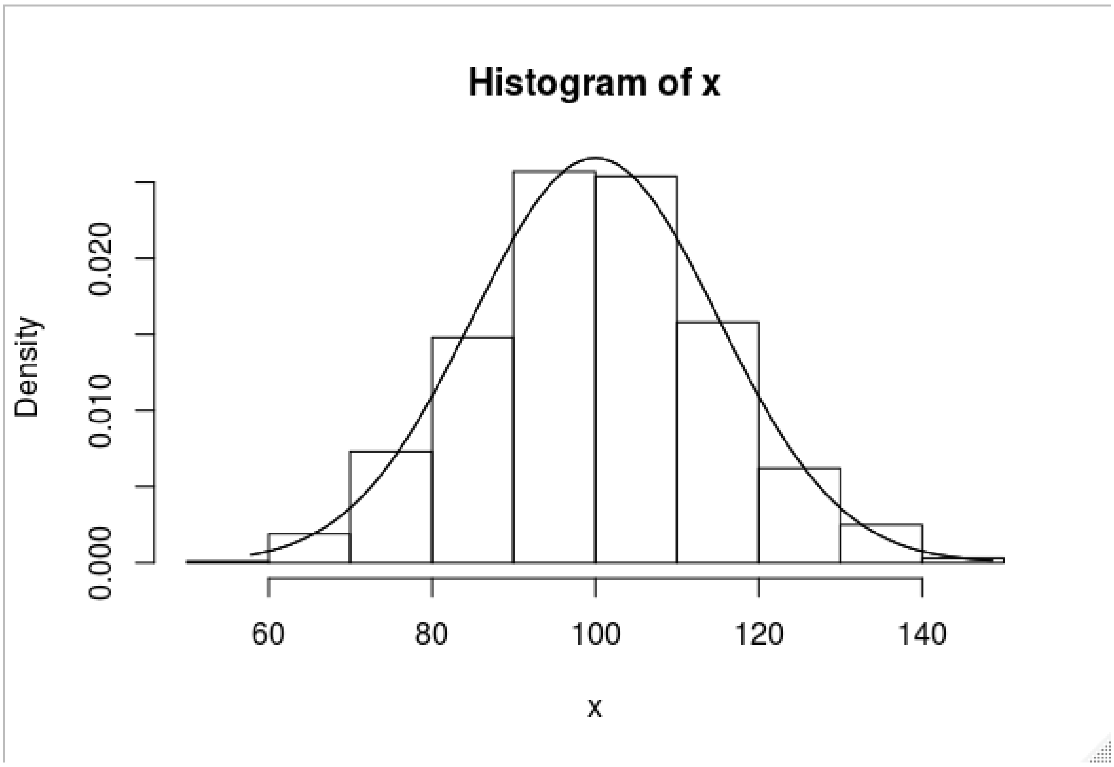
This plots the histogram (density function) of the above vector ‘x’

xx <- seq(min(x), max(x), length=100)

Here we make another vector which contains 100 values to plot the X-axis against density function (These values are in Arithmetic Progression with minimum and maximum specified)

lines(xx, dnorm(xx, mean=100, sd=15))

By this line we draw out a smooth curve of PDF using the points in xx(basically plot x vs y)



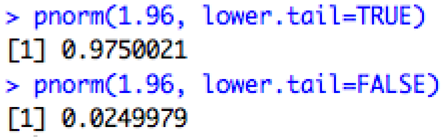
2. What is ***P*(*X* > 19)** when *X* has the **N(17.46, 375.67)** distribution?

This question asks the probability of X being greater than 19 given that X is a normal distribution with mean 17.46 and standard deviation 375.67

Function Used : pnorm()

Answer: 49.83646% chances having probability of X greater than 19.

3. Interpret the following



```
> pnorm(1.96, lower.tail=TRUE)
[1] 0.9750021
> pnorm(1.96, lower.tail=FALSE)
[1] 0.0249979
```

Interpretation: Here mean = 0 and sd = 1. So the first statement is probability of having a number lesser than 1.96. Second is having a probability greater than 1.96. Therefore their addition is 1(complimentary events)

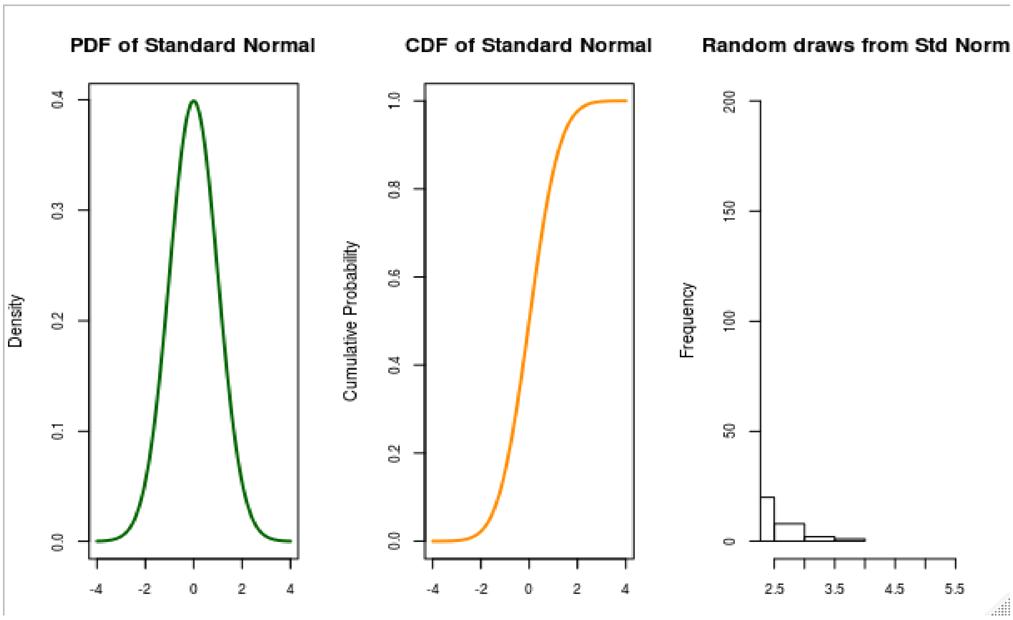
4. Run this in RStudio Script editor and explain it from plot

```
set.seed(3000)
xseq<- seq(-4,4,.01)
densities<- dnorm(xseq, 0,1)
cumulative<- pnorm(xseq, 0, 1)
randomdeviates<- rnorm(1000,0,1)
 par(mfrow=c(1,3), mar=c(3,4,4,2))

plot(xseq, densities, col="darkgreen",xlab="", ylab="Density", type="l",lwd=2, cex=2,
main="PDF of Standard Normal", cex.axis=.8)

plot(xseq, cumulative, col="darkorange", xlab="", ylab="Cumulative
Probability",type="l",lwd=2, cex=2, main="CDF of Standard Normal", cex.axis=.8)

hist(randomdeviates, main="Random draws from Std Normal", cex.axis=.8,
xlim=c(4,4))
```



Interpretation: The green is a graph of pdf it gives the density function of a Normal Function. The yellow is a graph of cdf it gives the cumulative density upto that value of X. (Therefore an increasing function). The frequency curve is the number of times some random number has occurred(Histogram Plot)

Conclusion:

- We could calculate the pdf and cdf of known distributions using R and therefore could verify our manual calculations.
- We understood the characteristics of various distributions.
- We could also plot pdf and cdf of various distributions which would take a lot of time to manually plot
- We could also get x% of some distribution which would take efforts to calculate manually
- We could associate real life events with probability distribution functions best suited and hence calculate and visualize the probability density function.