1)

2.A. Write a short note on Normal Distribution.

ANSWER:

- The most famous, and most used, statistical distribution is the normal distribution, sometimes referred to as the Gaussian distribution, which is defined as

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

where

$\mu =$ mean of x

$\sigma =$ standard deviation of x

$\pi = 3.14159...$

$e = 2.71828...$

Rnorm():

- To draw random numbers from the normal distribution use the rnorm function, which optionally allows the specification of the mean and standard deviation.

```
    >#10 draws from the standard 0-1 normal distribution
    >rnorm(n=10)
    [1]0.3746584 0.7368645 0.2408023 -0.1220292 0.6525665
    [6]0.3313728 0.5401996 1.6598050 -0.7777772 0.4904597

    >#10 draws from the 100-20 distribution
    >rnorm(n=10, mean=100, sd=20)
    [1] 94.99245 125.31772 120.70047 118.07148 111.88081 99.32752
[7] 92.36758 87.94429  115.18968 91.88554
```

Dnorm():

- The density for the normal distribution is calculated using dnorm.

```
>randNorm10<-rnorm(10)
randNorm10
[1] 1.8081780 0.7159731 0.4119520  -0.1659213 -0.1597631
[6] 1.0941883 0.1981299 -1.3998152 -2.2787374 -0.3403679
```

>dnorm(randNorm10)

```
[1] 0.07779389 0.30874263 0.36648754 0.39348848 0.39388328
[6] 0.21924564 0.39118830 0.14976620 0.02974005 0.37649004
```

>dnorm(c(-1,0,1))

```
[1] 0.2419707 0.3989423 0.2419707
```

- dnorm returns the probability of a specific number occurring. While it is technically mathematically impossible to find the exact probability of a number from a continuous distribution, this is an estimate of the probability. Like with rnorm, a mean and standard deviation can be specified for dnorm.

Pnorm():

```
>pnorm(randNorm10)
[1]0.96471060 0.76299601 0.65981269 0.43410943 0.43653383 0.86306380
[7]0.57852828 0.08078433 0.01134134 0.36678975
>pnorm(c(-3,0,3))
[1] 0.001349898 0.500000000 0.998650102
>pnorm(-1)
[1] 0.1586553
```

- By default this is left-tailed. To find the probability that the variable falls between two points, we must calculate the two probabilities and subtract them from each other.

```
>pnorm(1) – pnorm(0)
[1] 0.3413447
>pnorm(1)-pnorm(-1)
[1] 0.6826895
```

B. With examples explain in detail about Binomial distribution.

ANSWER:

- Like the normal distribution, the binomial distribution is well represented in R. Its probability mass function is

$$p(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n - x)!}$$

and n is the number of trails and p is the probability of success of a trail.

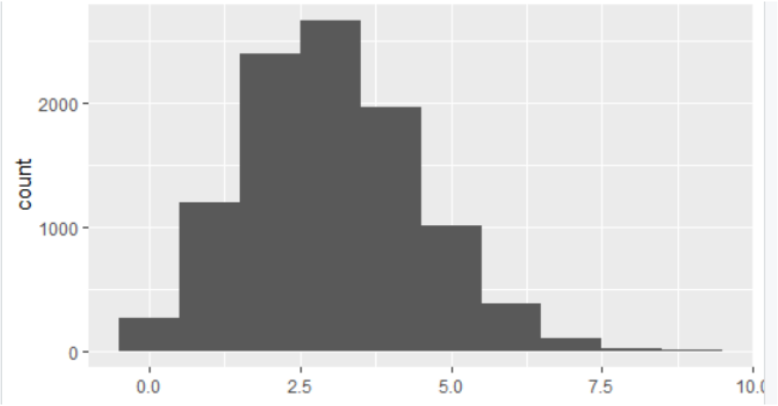
- The mean is np and the variance is np(1-p). When n=1 this reduces to the bernoulli distribution.
- Generating random numbers from the binomial distribution is not simply generating random numbers but rather generating the number of successes of independent trails.

- To simulate the number of successes out of ten trails with probability 0.4 of success, we run rbinom with n=1 (only one run of the trails), size=10 (trail size of 10), and prob=0.4 (probability of success is 0.4).

Rbinom():

```
>rbinom(n=1,size=10,prob=0.4)
[1]  6

>binomData<-data.frame(Successes=rbinom(n=10000,size=10,prob=0.3))

>ggplot(binomData, aes(x=Successes)) + geom_histogram(binwidth=1)
```

Output:



Dbinom():

```
>dbinom(x=3, size=10, prob=0.3)
[1]  0.2668279
```

Pbinom():

```
>#probability of 3 or fewer successes out of 10
>pbinom(q=3, size=10, prob=0.3)
[1]  0.6496107
```

Qbinom():

- Given a certain probability, qbinom returns the quantile, which for this distribution is the number of successes.

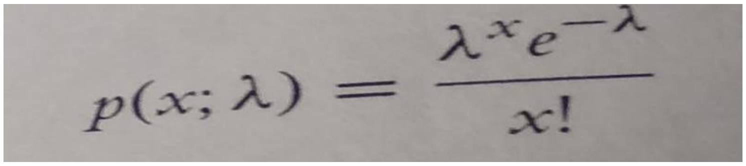
```
>qbinom(p=0.3,size=10,prob=0.3)
[1]  2
>qbinom(p=c(0.3,0.35,0.4,0.5,0.6), size=10,prob=0.3)
[1]  2  2  3  3  3
```

\

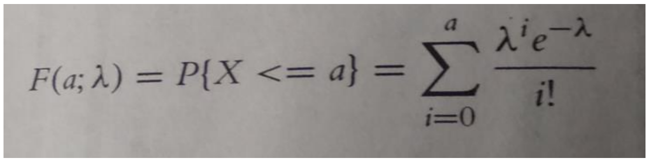
C. Write a short note on Poisson distribution.

ANSWER:

- Another popular distribution is the Poisson Distribution, which is for count data. Its probability mass function is



$$p(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$



$$F(a;\lambda) = P\{X <= a\} = \sum_{i=0}^{a} \frac{\lambda^i e^{-\lambda}}{i!}$$

Where λ is both the mean and variance.

- To generate random counts, the density, the distribution and quantiles use rpois, dpois, ppois and qpois, respectively.
- As λ grows large the poisson distribution begins to resemble the normal distribution. To see this we will simulate 10,000 draws from the Poisson distribution and plot their histograms to see the shape.

```
>#generate 10,000 random counts from 5 different Poisson distributions
>pois1<-rpois(n=10000, lambda=1)
>pois2<-rpois(n=10000, lambda=2)
>pois5<-rpois(n=10000, lambda=5)
>pois10<-rpois(n=10000,lambda=10)
>pois20<-rpois(n=10000,lambda=20)
>pois<-data.frame(Lambda.1=pois1, Lambda.2=pois2, Lambda.5=pois5, Lambda.10=pois10,
Lambda.20=pois20)
```

----

3. A. Write a short note on summary() with examples.

ANSWER:

- The first thing many people think of in relation to statistics is the average, or mean, as it is properly called.

```
>x<-sample(x=1:100, size=100, replace= TRUE)
>x
[1]    81 100 79 66 32 87 85 10 46 76 71 25 100 51
[15] 17 60 15 6 60 89 23 100 64 28 74 12 12 10
[29] 48 79 10 87 50 21 96 85 40 41 64 18 92 70
[43] 44 28 73 22 35 93 18 64 43 50 53 52 84 99
```

```
[57] 25 61 41 47 1 66 93 22 77 56 17 74 19 85
[71] 25 11 27 83 100 37 53 64 60 55 69 58 62 53
[85] 62 36 96 68 11 54 45 69 35 63 38 16 95 1
[99] 80 69
```

mean(x)

```
[1] 53.17
```

- This is the simple arithmetic mean

$$E[X] = \frac{\sum_{i=1}^{N} x_i}{N}$$

```
mean(y,na.rm=TRUE)
[1] 51.025
```

Weighted Mean

```
weighted.mean(x=grades,w=weights)
[1] 84.625
```

- The formula for weighted.mean is given by

$$E[X] = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \sum_{i=1}^{N} p_i x_i$$

Variance:

```
var(x)
[1] 710.0469
```

- This calculates variance as

$$Var(x) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}$$

Standard Deviation:

```
sd(x)
[1] 26.6467
>sd(y,na.rm= TRUE)   #if u want to discard NA values and then calculate
[1] 27.4761
```

- Other commonly used functions summary statistics are **min, max and median**. Of course all of these also have na.rm arguments.

```
>min(x)
[1]  3
>max(x)
[1]  99
>median(x)
[1]  43
>min(y)
```

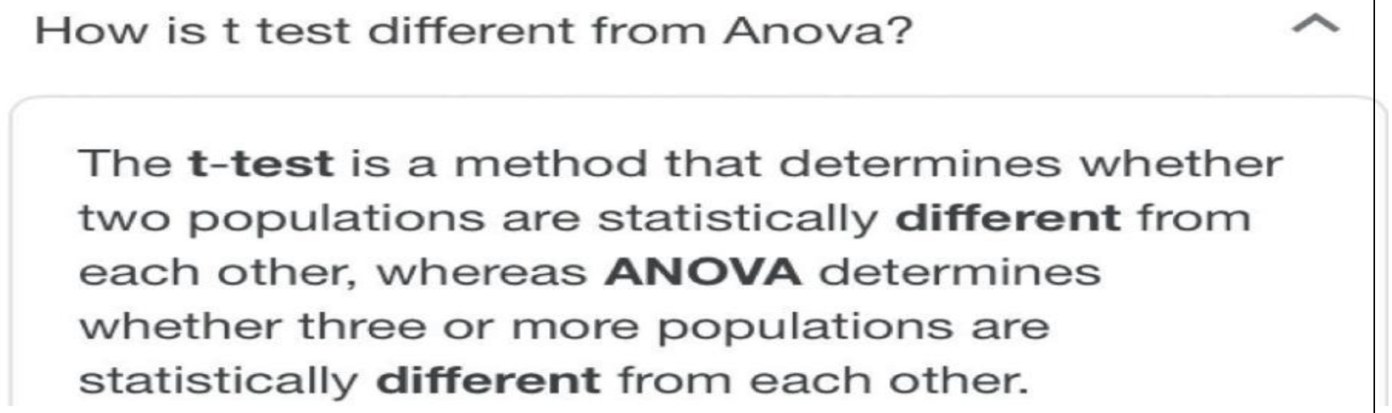
```
[1]  NA      #because y may contain NA values
```

Summary:

```
>summary(x)
Min     1st Qu. Median          Mean   3rd Qu. Max
1.00    17.75   43.00   44.51   68.25   100.00
>summary(y)
Min     1st Qu. Median          Mean   3rd Qu. Max     NA's
2.00    18.00   40.50   43.59   67.00   100.00 20
```

6.Write a short note on ANOVA.

ANSWER:



How is t test different from Anova?

The **t-test** is a method that determines whether two populations are statistically **different** from each other, whereas **ANOVA** determines whether three or more populations are statistically **different** from each other.

$$F = \frac{\sum_{i} n_i(\bar{Y}_i - \bar{Y})^2/(K-1)}{\sum_{ij}(Y_{ij} - \bar{Y}_i)^2/(N-K)}$$

R-Code:

```
>tipAnova<-aov(tip~day-1,tips)
>tipIntercept<-aov(tip~day,tips)
>tipAnova$coefficients
```

Output:

```
        dayFri      daySat              daySun              dayThur

         2.734737 2.993103 3.255132   2.771452
```

```
>tipIntercept$coefficients
```

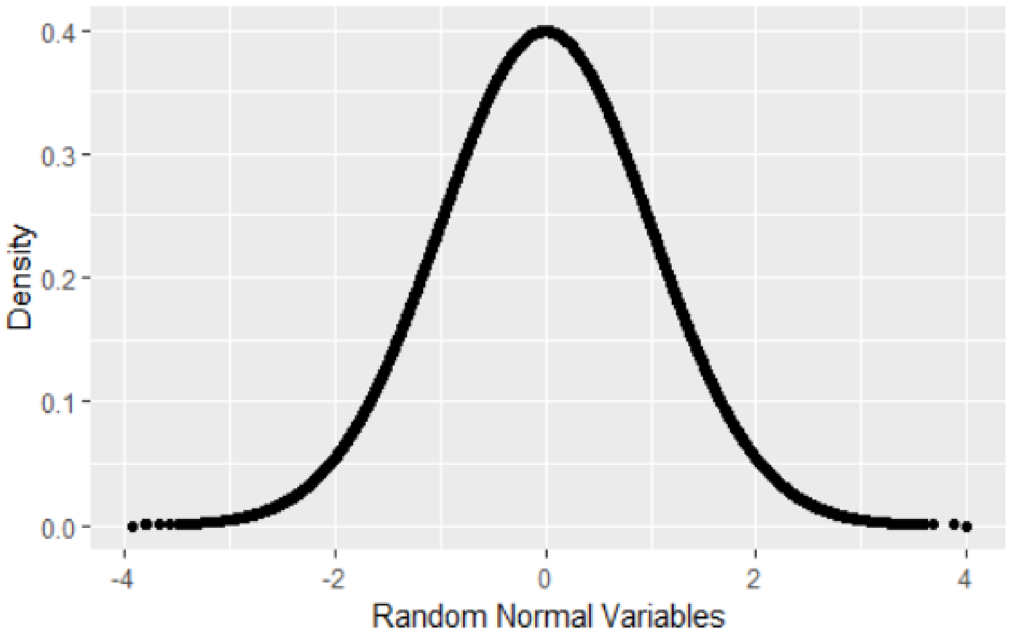
Difference between correlation and covariance:

Covariance	**Correlation**
Covariance is a measure to indicate the extent to which two random variables change in tandem.	Correlation is a measure used to represent how strongly two random variables are related to each other.
Covariance is nothing but a measure of correlation.	Correlation refers to the scaled form of covariance.
Covariance indicates the direction of the linear relationship between variables.	Correlation on the other hand measures both the strength and direction of the linear relationship between two variables.
Covariance can vary between -∞ and +∞	Correlation ranges between -1 and +1
Covariance is affected by the change in scale. If all the values of one variable are multiplied by a constant and all the values of another variable are multiplied, by a similar or different constant, then the covariance is changed.	Correlation is not influenced by the change in scale.
Covariance assumes the units from the product of the units of the two variables.	Correlation is dimensionless, i.e. It’s a unit-free measure of the relationship between variables.
Covariance of two dependent variables measures how much in real quantity (i.e. cm, kg, liters) on average they co-vary.	Correlation of two dependent variables measures the proportion of how much on average these variables vary w.r.t one another.

Covariance is zero in case of independent variables (if one variable moves and the other doesn’t) because then the variables do not necessarily move together.	Independent movements do not contribute to the total correlation. Therefore, completely independent variables have a zero correlation.

Write the code to plot “random normal variables and their densities”, which results in a bell curve

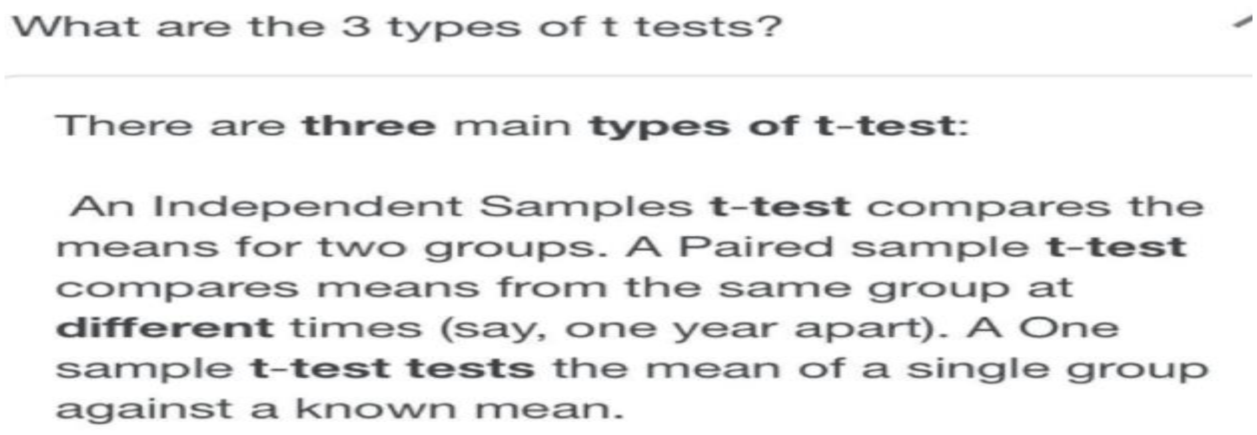
```
>#generate the normal variables

>randNorm<rnorm(30000)

>#calculate their distributions

>randDensity<-dnorm(randNorm)

>#load ggplot2

>require(ggplot2)

>#plot them

>ggplot(data.frame(x=randNorm, y=randDensity)) + aes(x=x,y=y) + geom_point() + labs(x=“Random Normal Variables”, y= “Density”)
```



Write a short note on one-sample test, two-sample t-test and paired sample two-test.

One –sample t test

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.



What are the 3 types of t tests?

There are **three** main **types of t-test**:

An Independent Samples **t-test** compares the means for two groups. A Paired sample **t-test** compares means from the same group at **different** times (say, one year apart). A One sample **t-test tests** the mean of a single group against a known mean.

One Sample T-test:

R-code:

```
            t.test(tips$tip,alternative=“two.sided”, mu=2.5)
```

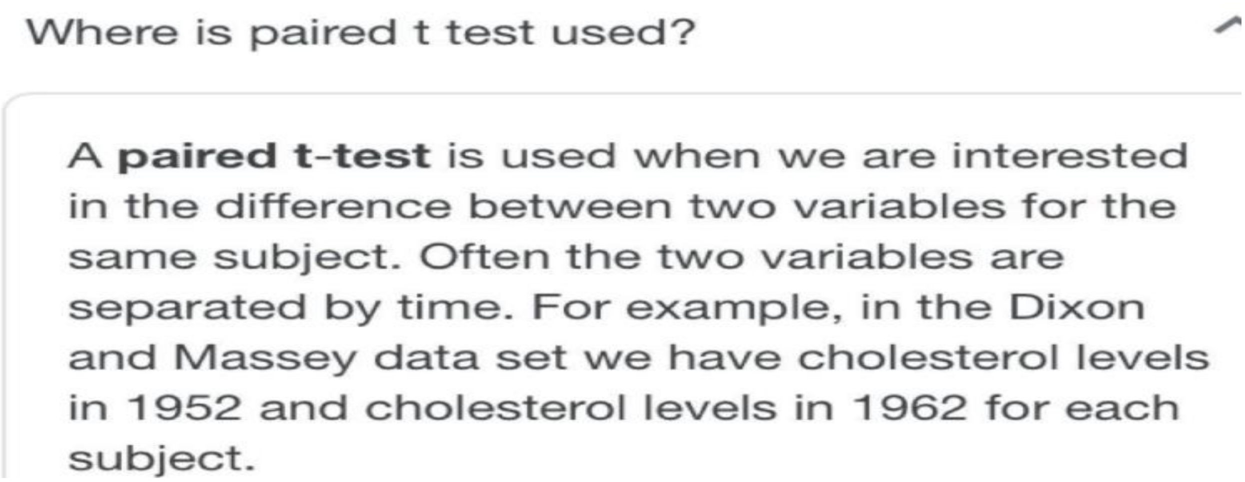
Two-Sample T-Test:

- More often or not the t-test is used for comparing two samples.
- Continuing with the tips data, we compare how female and male servers are tipped. Before running the t-test, however, we first need to check the variance, whereas the Welch two-sample t-test can handle groups with differing variances.

```
>#first just compute the variance for each group;

>#using the formula interface

>#calculate the variance of tip for each level of sex

>aggregate(tip~sex,data=tips,var)

           sex  tip

    1   Female     1.344428

    2   Male       2.217424

>#now test for normality of tip distribution

>shapiro.test(tips$tip)
```

Output:

```
Shapiro-Wilk normality test

data:  tips$tip

W = 0.89781, p-value = 8.2e-12
```



Where is paired t test used?

A **paired t-test** is used when we are interested in the difference between two variables for the same subject. Often the two variables are separated by time. For example, in the Dixon and Massey data set we have cholesterol levels in 1952 and cholesterol levels in 1962 for each subject.

- For testing paired data a paired t-test should be used. This is simple enough to do by setting the paired argument in t.test to TRUE.

- Heights are generally normally distributed, so we will forgo the tests of normality and equal variance.

- **Paired sample two-test:**
- **R-code:**
  - t.test(father.son$fheight, father.son$sheight, paired=TRUE)