

UNIT-V: Probability Distributions, Normal Distribution- Binomial Distribution- Poisson Distributions, Other Distribution, Basic Statistics, Correlation and Covariance, T-Tests, ANOVA.

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.

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

Annotations for the formula:

- $n!$: This starts the count of number of ways event can occur.
- $(n-x)!$: This ends the count of number of ways event can occur.
- $x!$: This deletes duplications.
- p^x : This is the probability of success for x trials.
- q^{n-x} : This is the probability of failure for the x trials.

$$\text{Mean} = \mu = E(x) = np$$

$$\text{Variance} = \sigma^2 = np(1-p)$$

$$\text{Standard Deviation} = \sigma = \sqrt{np(1-p)}$$

where

n = number of trials

p = probability of success

$1-p$ = probability of other outcome (failure)

R has four in-built functions to generate binomial distribution. They are described below.

- ***dbinom(x, size, prob)*** :- This function gives the probability density distribution at each point.
- ***pbinom(x, size, prob)*** :- This function gives the cumulative probability of an event. It is a single value representing the probability.
- ***qbinom(p, size, prob)*** :- This function takes the probability value and gives a number whose cumulative value matches the probability value.
- ***rbinom(n, size, prob)*** :- This function generates required number of random values of given probability from a given sample.

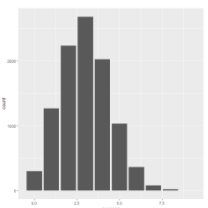
Following is the description of the parameters used –

- ✓ x is a vector of numbers.
- ✓ p is a vector of probabilities.
- ✓ n is number of observations.
- ✓ $size$ is the number of trials.
- ✓ $prob$ is the probability of success of each trial.

Examples:

- ***rbinom(n=1,size=10,prob=0.4)*** - It generates 1 random number from the binomial distribution based on number of successes of 10 independent trials.
- ***rbinom(n=5,size=10,prob=0.4)*** - It generates 5 random number from the binomial distribution based on number of successes of 10 independent trials with probability 0.4.
- ***rbinom(n=5,size=1,prob=0.4)*** - Setting size to 1 turns the numbers into a bernoulli random variable, which can take only value 1 (success) or 0 (failure).
- To visualize the binomial distribution we randomly generate 10,000 experiments, each with 10 trials and 0.3 probability.

```
b <- data.frame(success=rbinom(n=10000,size=10,prob=0.3))
ggplot(b,aes(x=success))+geom_bar()
```



Problem: Suppose a die is tossed 5 times. What is the probability of getting exactly 2 fours?

Solution: This is a binomial experiment in which the number of trials is equal to 5, the number of successes is equal to 2, and the probability of success on a single trial is $1/6$ or about 0.167. Therefore, the binomial probability is:

$$b(2; 5, 0.167) = {}^5C_2 * (0.167)^2 * (0.833)^3$$

$$b(2; 5, 0.167) = 0.161$$

R Code:

```
> dbinom(2, size=5, prob=0.167)
[1] 0.1612
```



Problem: In a restaurant seventy percent of people order for Chinese food and thirty percent for Italian food. A group of three persons enter the restaurant. Find the probability of at least two of them ordering for Italian food.

Solution:-

The probability of ordering Chinese food is 0.7 and the probability of ordering Italian food is 0.3. Now, if at least two of them are ordering Italian food then it implies that either two or three will order Italian food.

Probability for two ordering Italian food,

$$\begin{aligned} P(X=2) &= {}^3C_2(0.3)^2(0.7)^1 \\ &= 3 \times 0.09 \times 0.7 \\ &= 0.189 \end{aligned}$$

Probability for all three ordering Italian food,

$$\begin{aligned} P(X=3) &= {}^3C_3(0.3)^3(0.7)^0 \\ &= 1 \times 0.027 \times 1 \\ &= 0.027 \end{aligned}$$

Hence, the probability for at least two persons ordering Italian food is,

$$P(X \geq 2) = P(X=2) + P(X=3) = 0.189 + 0.027 = 0.216$$

R code:-

```
> dbinom(2,size=3,prob=0.3)+
+ dbinom(3,size=3,prob=0.3)
[1] 0.216
```



Cumulative Binomial Probability:- A cumulative binomial probability refers to the probability that the binomial random variable falls within a specified range (e.g., is greater than or equal to a stated lower limit and less than or equal to a stated upper limit).

Problem: What is the probability of obtaining 45 or fewer heads in 100 tosses of a coin?

Solution: To solve this problem, we compute 46 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek.

Thus,

$$\begin{aligned} b(x \leq 45; 100, 0.5) &= b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + \dots + b(x = 45; 100, 0.5) \\ &= 0.184 \end{aligned}$$

R code:-

```
> pbinom(45,size=100,prob=0.5)
[1] 0.1841008
```



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.

Solution:

Since only one out of five possible answers is correct, the probability of answering a question correctly by random is $1/5=0.2$.

- To 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 $x = 0, \dots, 4$.

```
> 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%.

Problem: Fit an appropriate binomial distribution and calculate the theoretical distribution

```
x: 0 1 2 3 4 5
f: 2 14 20 34 22 8
```

Solution:

Here $n = 5$, $N = 100$
 Mean = $\frac{\sum x_i f_i}{\sum f_i} = 2.84$
 $np = 2.84$
 $p = 2.84/5 = 0.568$
 $q = 0.432$

$p(r) = {}^5C_r (0.568)^r (0.432)^{5-r}$, $r = 0, 1, 2, 3, 4, 5$

Theoretical distributions are

Calculation of Expected Frequency as follows

r	p(r)	N* p(r)
0	0.0147	100 * 0.0147 = 1.47 = 1
1	0.097	100 * 0.097 = 9.7 = 10
2	0.258	100 * 0.258 = 25.8 = 26
3	0.342	100 * 0.342 = 34.2 = 34
4	0.226	100 * 0.226 = 22.6 = 23
5	0.060	100 * 0.060 = 6 = 6
		Total = 100

R code:-

```
> x <- 0:5
> f <- c(2,14,20,34,22,8)
> df <- data.frame(x,f)
> fitbin <- fitdist(df$f,"nbinom")
> summary(fitbin)
```

Fitting of the distribution 'nbinom' by maximum likelihood

Parameters :

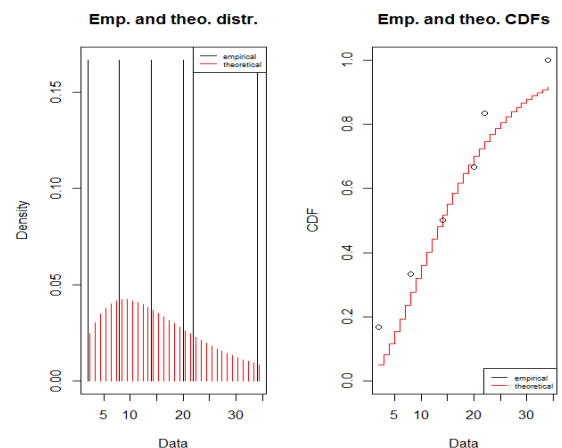
	estimate	Std. Error
size	2.192416	1.441296
mu	16.664004	4.886713

Loglikelihood: -22.387 AIC: 48.774 BIC: 48.35752

Correlation matrix:

	size	mu
size	1.0000000000	0.0003165092
mu	0.0003165092	1.0000000000

```
> plot(fitbin)
```



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:

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

$p(x)$ = Probability of x given λ
 λ = Expected (mean) number 'successes'
 e = 2.71828 (base of natural logs)
 x = Number of 'successes' in per unit

Mean **Standard Deviation**
 $\mu = E(x) = \lambda$ $\sigma = \sqrt{\lambda}$

Examples:

1. The number of defective electric bulbs manufactured by a reputed company.
2. The number of telephone calls per minute at a switch board
3. The number of cars passing a certain point in one minute.
4. The number of printing mistakes per page in a large text.

R has four in-built functions to generate binomial distribution. They are described below.

- **dpois(x, lambda, log = FALSE)** :- This function gives the probability density distribution at each point.
- **ppois(q, lambda, lower.tail = TRUE, log.p = FALSE)** :- This function gives the cumulative probability of an event. It is a single value representing the probability.
- **qpois(p, lambda, lower.tail = TRUE, log.p = FALSE)** :- This function takes the probability value and gives a number whose cumulative value matches the probability value.
- **rpois(n, lambda)** :- This function generates required number of random values of given probability from a given sample.

Following is the description of the parameters used –

- ✓ x is a vector of numbers.
- ✓ p is a vector of probabilities.
- ✓ n is number of observations.
- ✓ size is the number of trials.
- ✓ prob is the probability of success of each trial.

 **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.

Solution:-

The probability of having sixteen or less cars crossing the bridge in a particular minute is given by the function ppois.


```
> ppois(16, lambda=12) # lower tail
[1] 0.89871
```

Hence the probability of having seventeen or more cars crossing the bridge in a minute is in the upper tail of the probability density function.

```
> ppois(16, lambda=12, lower=FALSE) # upper tail
[1] 0.10129
```

Answer:- If there are twelve cars crossing a bridge per minute on average, the probability of having seventeen or more cars crossing the bridge in a particular minute is 10.1%.



 **Problem:-** The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution: This is a Poisson experiment in which we know the following:

- $\mu = 2$; since 2 homes are sold per day, on average.
- $x = 3$; since we want to find the likelihood that 3 homes will be sold tomorrow.
- $e = 2.71828$; since e is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:



$$\begin{aligned}
 P(x; \mu) &= (e^{-\mu}) (\mu^x) / x! \\
 P(3; 2) &= (2.71828^{-2}) (2^3) / 3! \\
 &= (0.13534) (8) / 6 \\
 &= 0.180
 \end{aligned}$$

R Code:-

```
> dpois(3, lambda = 2)
[1] 0.180447
```

Cumulative Poisson Probability:- A **cumulative Poisson probability** refers to the probability that the Poisson random variable is greater than some specified lower limit and less than some specified upper limit.

Problem:- Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

Solution: This is a Poisson experiment in which we know the following:

- $\mu = 5$; since 5 lions are seen per safari, on average.
- $x = 0, 1, 2, \text{ or } 3$; since we want to find the likelihood that tourists will see fewer than 4 lions; that is, we want the probability that they will see 0, 1, 2, or 3 lions.
- $e = 2.71828$; since e is a constant equal to approximately 2.71828.



To solve this problem, we need to find the probability that tourists will see 0, 1, 2, or 3 lions. Thus, we need to calculate the sum of four probabilities: $P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)$. To compute this sum, we use the Poisson formula:

$$P(x \leq 3, 5) = P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)$$

$$P(x \leq 3, 5) = [(e^{-5})(5^0) / 0!] + [(e^{-5})(5^1) / 1!] + [(e^{-5})(5^2) / 2!] + [(e^{-5})(5^3) / 3!]$$

$$P(x \leq 3, 5) = [(0.006738)(1) / 1] + [(0.006738)(5) / 1] + [(0.006738)(25) / 2] + [(0.006738)(125) / 6]$$

$$P(x \leq 3, 5) = [0.0067] + [0.03369] + [0.084224] + [0.140375]$$

$$P(x \leq 3, 5) = 0.2650$$

Thus, the probability of seeing at no more than 3 lions is 0.2650.

R Code:-

```
> ppois(3, lambda = 5)
[1] 0.2650259
```

Normal Distribution:- A continuous random variable X follows a normal distribution with mean μ and variance σ^2 is a statistic distribution with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\mu-x)^2}{2\sigma^2}}, \quad \text{on the domain } x \in (-\infty, \infty).$$

Standard Normal Distribution

It is the distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one.

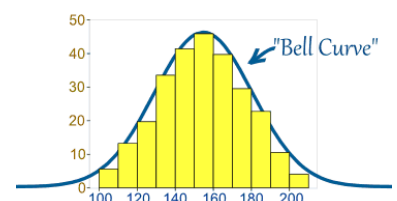
The normal random variable of a standard normal distribution is called a standard score or a z score. Every normal random variable X can be transformed into a z score via the following equation:

$$Z = (X - \mu) / \sigma$$

where X is a normal random variable, μ is the mean, and σ is the standard deviation. yielding

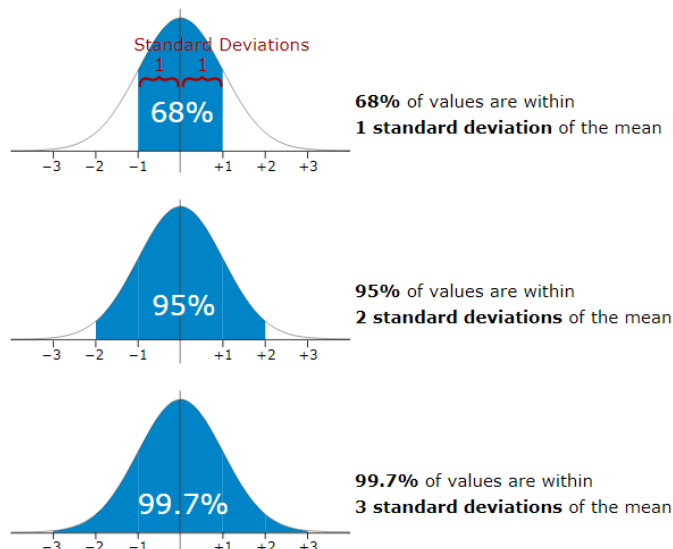
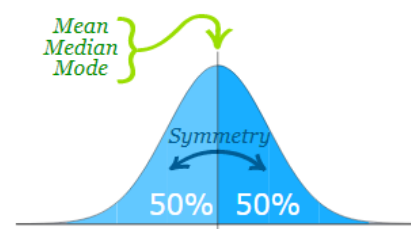
$$P(x) dx = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Standard Normal Curve:- One way of figuring out how data are distributed is to plot them in a graph. If the data is evenly distributed, you may come up with a bell curve. A bell curve has a small percentage of the points on both tails and the bigger percentage on the inner part of



the curve. The shape of the standard normal distribution looks like this:

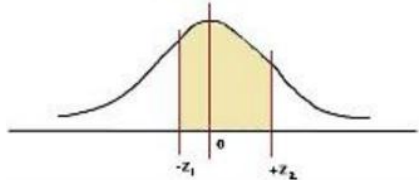
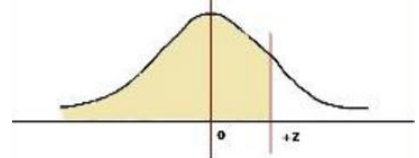
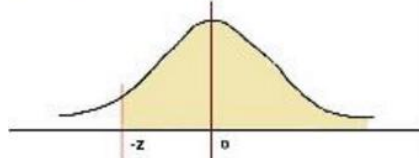
- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean



R functions:

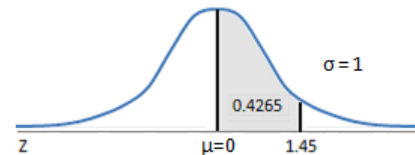
- `dnorm(x, mean = 0, sd = 1, log = FALSE)` :- This function gives the probability density distribution at each point.
- `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`:- This function gives the cumulative probability of an event. It is a single value representing the probability.
- `qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`:- This function takes the probability value and gives a number whose cumulative value matches the probability value.
- `rnorm(n, mean = 0, sd = 1)` :- This function generates required number of random values of given probability from a given sample.

Procedure to find probability using positive Z-score table			
Case 1: Area between 0 and any z score	Area(z)		
Case 2: Area in any tail	0.5 - Area(z)		
Case 3: Area between two z-scores on the same side of the mean	Area(z2)-Area(z1)		

Case 4: Area between two z-scores on the opposite side of the mean	$\text{Area}(z_1) + \text{Area}(z_2)$	
Case 5: Area to the left of a positive Z score	$0.5 + \text{Area}(z)$	
Case 6: Area to the right of a negative Z score	$0.5 + \text{Area}(z)$	

Areas Under the One-Tailed Standard Normal Curve

This table provides the area between the mean and some Z score.
For example, when Z score = 1.45
the area = 0.4265.



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

- Problem:-** X is a normally distributed variable with mean $\mu = 30$ and standard deviation $\sigma = 4$. Find
- $P(x < 40)$
 - $P(x > 21)$
 - $P(30 < x < 35)$

Solution:

- For $x = 40$, then

$$z = \frac{x - \mu}{\sigma}$$

$$\Rightarrow z = \frac{40 - 30}{4}$$

$$= 2.5 \text{ (= } z_1 \text{ say)}$$
Hence $P(x < 40) = P(z < 2.5)$

$$= 0.5 + A(z_1) = 0.9938$$
- For $x = 21$,

$$z = \frac{x - \mu}{\sigma}$$

$$\Rightarrow z = \frac{21 - 30}{4}$$

$$= -2.25 \text{ (= } -z_1 \text{ say)}$$
Hence $P(x > 21) = P(z > -2.25)$

$$= 0.5 - A(z_1) = 0.9878$$
- For $x = 30$

$$z = \frac{x - \mu}{\sigma} \Rightarrow,$$

$$z = \frac{30 - 30}{4} = 0 \text{ and}$$
for $x = 35$,

$$z = \frac{x - \mu}{\sigma}$$

$$\Rightarrow z = \frac{35 - 30}{4}$$

$$= 1.25$$
Hence $P(30 < x < 35) = P(0 < z < 1.25)$

$$= [\text{area to the left of } z = 1.25] - [\text{area to the left of } 0]$$

$$= 0.8944 - 0.5 = 0.3944$$

Problem:- The length of life of an instrument produced by a machine has a normal distribution with a mean of 12 months and standard deviation of 2 months. Find the probability that an instrument produced by this machine will last.

- less than 7 months.
- between 7 and 12 months.

Solution:

- $P(x < 7)$
for $x = 7$

$$z = \frac{x - \mu}{\sigma}$$

$$\Rightarrow z = \frac{7 - 12}{2}$$

$$= -2.5 \text{ (= } z_1 \text{ say)}$$
Hence $P(x < 7) = P(z < -2.5)$

$$= 0.0062$$
- $P(7 < x < 12)$
For $x=12$

$$z = \frac{x - \mu}{\sigma}$$

$$\Rightarrow z = \frac{12 - 12}{2}$$

$$= 0 \text{ (= } z_1 \text{ say)}$$
Hence $P(7 < x < 12) = P(-2.5 < z < 0)$

$$= 0.4938$$



Problem:- The Tahoe Natural Coffee Shop morning customer load follows a normal distribution with mean 45 and standard deviation 8. Determine the probability that the number of customers tomorrow will be less than 42.



Solution:-

We first convert the raw score to a z-score. We have

$$z = \frac{x - \mu}{\sigma}$$

$$\Rightarrow z = (42 - 45) / 8 = -0.375$$

Next, we use the table to find the probability. The table gives 0.3520. (We have rounded the raw score to -0.38).

We can conclude that

$$P(x < 42) = P(z < -0.38)$$

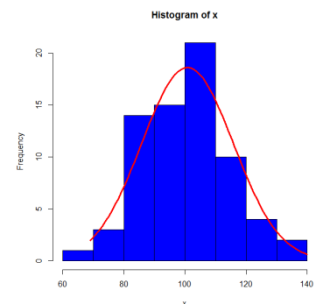
$$= 0.352$$

That is there is about a 35% chance that there will be fewer than 42 customers tomorrow.

Example:

```
> x <- c(92,117,109,85,117,107,82,83,119,113,101,106,101,84,126,69,82,79,84,100,104,111,109,92,93,107,
81,118,81,133,111,82,120,103,115,89,74,110,83,110,96,102,108,110,140,106,111,98,98,99,74,101,107,104,
128,87,95,109,104,91,83,98,99,103,126,123,85,98,93,100)
```

```
> h<-hist(x,col = "blue")
> m <- mean(x)
> s <- sd(x)
> xf <- seq(min(x),max(x),length=70)
> dis <- dnorm(xf,m,s)
> dis <- dis*diff(h$mids[1:2]*length(x))
> lines(xf,dis,col="red",lwd=3)
```



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?

Solution:-

We apply the function pnorm of the normal distribution with mean 72 and standard deviation 15.2. Since we are looking for the percentage of students scoring higher than 84, we are interested in the *upper tail* of the normal distribution.

```
> pnorm(84, mean=72, sd=15.2, lower.tail=FALSE)
[1] 0.21492
```

Correlation:- A correlation is a relationship between two variables. Typically, we take x to be the independent variable. We take y to be the dependent variable. Data is represented by a collection of ordered pairs (x,y).

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

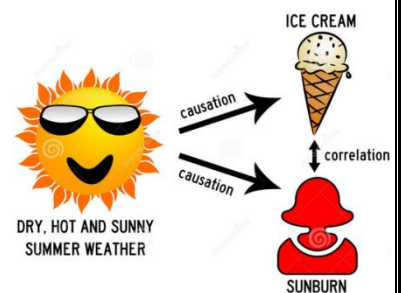
This will always be a number between -1 and 1 (inclusive).

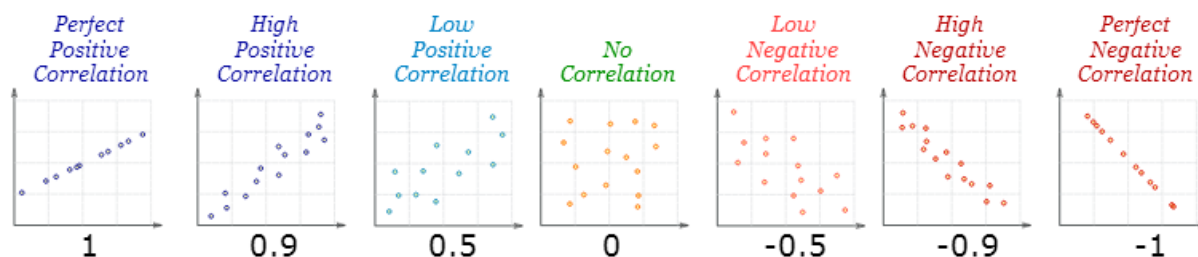
- If r is close to 1, we say that the variables are positively correlated. This means there is likely a strong linear relationship between the two variables, with a positive slope.
- If r is close to -1, we say that the variables are negatively correlated. This means there is likely a strong linear relationship between the two variables, with a negative slope.
- If r is close to 0, we say that the variables are not correlated. This means that there is likely no linear relationship between the two variables, however, the variables may still be related in some other way.

To run a correlation test we type:

```
> cor.test(var1, var2, method = "method")
```

The default method is "pearson" so you may omit this if that is what you want. If you type "kendall" or "spearman" then you will get the appropriate significance test.





Problem:- The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days:

Temperature oC	14.2	16.4	11.9	15.2	18.5	22.1	19.4	25.1	23.4	18.1	22.6	17.2
Ice cream sales	\$215	\$325	\$185	\$332	\$406	\$522	\$412	\$614	\$544	\$421	\$445	\$408

Solution:-

Formula for correlation coefficient:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

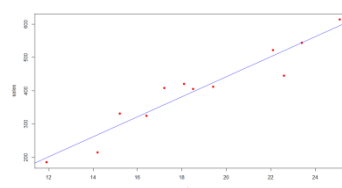


		2 Subtract Mean		3 Calculate ab, a² and b²		
Temp °C	Sales	"a"	"b"	a×b	a ²	b ²
14.2	\$215	-4.5	-\$187	842	20.3	34,969
16.4	\$325	-2.3	-\$77	177	5.3	5,929
11.9	\$185	-6.8	-\$217	1,476	46.2	47,089
15.2	\$332	-3.5	-\$70	245	12.3	4,900
18.5	\$406	-0.2	\$4	-1	0.0	16
22.1	\$522	3.4	\$120	408	11.6	14,400
19.4	\$412	0.7	\$10	7	0.5	100
25.1	\$614	6.4	\$212	1,357	41.0	44,944
23.4	\$544	4.7	\$142	667	22.1	20,164
18.1	\$421	-0.6	\$19	-11	0.4	361
22.6	\$445	3.9	\$43	168	15.2	1,849
17.2	\$408	-1.5	\$6	-9	2.3	36
18.7	\$402			5,325	177.0	174,757

5 $\frac{5,325}{\sqrt{177.0 \times 174,757}} = 0.9575$

R Code:-

```
> temp <- c(14.2,16.4,11.9,15.2,18.5,22.1,19.4,25.1,23.4,18.1,22.6,17.2)
> sales <- c(215,325,185,332,406,522,412,614,544,421,445,408)
> corr_coef <- cor(temp,sales)
> corr_coef
[1] 0.9575066
> cov(temp,sales)
[1] 484.0932
#Adds a line of best fit to your scatter plot
> plot(temp, sales, pch=16,col="red")
> abline(lm(sales~temp),col="blue")
```



Type I : This method is used when given variables are small in magnitude.

$$\text{Formula : } r = \frac{N \sum XY - \sum X \sum Y}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}}$$

Example 1. Calculate Karl Pearson's coefficient of correlation between the age and weight of the children :

Age (years) :	1	2	3	4	5
Weight (kg.) :	3	4	6	7	12

Solution : $\sum X = 15$; $\sum Y = 32$; $\sum X^2 = 55$; $\sum Y^2 = 254$; $\sum XY = 117$

Age (X)	Weight (Y)	X^2	Y^2	XY
1	3	1	9	3
2	4	4	16	8
3	6	9	36	18
4	7	16	49	28
5	12	29	144	60
15	32	55	254	117

$$\text{As } r = \frac{N \sum XY - \sum X \sum Y}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}}$$

$$\therefore r = \frac{5 \times 117 - 15 \times 32}{\sqrt{5 \times 55 - (15)^2} \sqrt{5 \times 254 - (32)^2}}$$

$$= \frac{585 - 480}{\sqrt{275 - 225} \sqrt{1270 - 1024}} = \frac{105}{\sqrt{50 \times 246}} = \frac{105}{\sqrt{12300}} = \frac{105}{110.90} = 0.9467 \text{ Ans.}$$

T-test for single mean:- One-sample t-test is used to compare the mean of a population to a specified theoretical mean (μ).

Let X represents a set of values with size n, with mean μ and with standard deviation S. The comparison of the observed mean (μ) of the population to a theoretical value μ is performed with the formula below:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

To evaluate whether the difference is statistically significant, you first have to read in t test table the critical value of Student's t distribution corresponding to the significance level alpha of your choice (5%). The degrees of freedom (df) used in this test are: $df = n-1$

Problem:- A professor wants to know if her introductory statistics class has a good grasp of basic math. Six students are chosen at random from the class and given a math proficiency test. The professor wants the class to be able to score above 70 on the test. The six students get scores of 62, 92, 75, 68, 83, and 95. Can the professor have 90 percent confidence that the mean score for the class on the test would be above 70?

Solution:-

Null hypothesis: $H_0: \mu = 70$

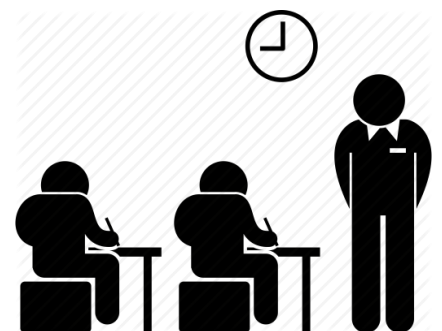
Alternative hypothesis: $H_a: \mu > 70$

First, compute the sample mean and standard deviation:

$$\bar{x} = \frac{62 + 92 + 75 + 68 + 83 + 95}{6}$$

$$= \frac{475}{6} = 13.17$$

- **Null Hypothesis** H_0 : The sample meet upto standard i.e $\mu > 70$ hours
- **Alternative Hypothesis** H_A : μ not greater than 70,
- **Level of Significance:** $\alpha = 0.05$
- **The test statistic is** $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$



$$t = \frac{79.71 - 70}{13.17/\sqrt{6}} = \frac{9.17}{5.38}$$

$$= 1.71 (\text{calculate value of } t)$$

To test the hypothesis, the computed t -value of 1.71 will be compared to the critical value in the t -table with 5 df is 1.67, the calculate of t is more than table value of t , so null hypothesis is rejected.

R code:-

```
> t.test(x, alternative="two.sided", mu=70)
```

One Sample t-test

```
data: x
t = 1.7053, df = 5, p-value = 0.1489
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval:
 65.34888 92.98446
sample estimates:
mean of x
 79.16667
```

Problem:- A Sample of 26 bulbs gives a mean life of 990 hours with S.D of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is sample meet upto the standard.

Solution: Here $n = 26$,

Sample mean $\bar{x} = 990$ hours

S.D $s = 20$ hours

Population mean $\mu = 1000$ hours

Df = $n-1 = 26-1 = 25$

- **Null Hypothesis H_0 :** The sample meet upto standard i.e $\mu = 1000$ hours
- **Alternative Hypothesis H_A :** μ not equal to 1000,
- **Level of Significance:** $\alpha = 0.05$
- **the test statistic is**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$t = 990 - 1000 / 20 / \sqrt{26}$$

$$= 2.5 (\text{calculate value of } t)$$

Table value of t with 25 df is 1.708

The calculate value of t is more than table value of t , so null hypothesis is rejected at 5% level.



Paired comparisons(Paired t-test):- Sometimes data comes from non independent samples. An example might be testing "before and after" of cosmetics or consumer products. We could use a single random sample and do "before and after" tests on each person. A hypothesis test based on these data would be called a *paired comparisons test*. Since the observations come in pairs, we can study the difference, d , between the samples. The difference between each pair of measurements is called d_i .

Test statistic:- With a population of n pairs of measurements, forming a simple random sample from a normally distributed population, the mean of the difference, \bar{d} , is tested using the following implementation of t .

$$t = \frac{\bar{d} - \mu}{S/\sqrt{n}}$$

Problem :- The blood pressure of 5 women before and after intake of a certain drug are given below: Test whether there is significant change in blood pressure at 1% level of significance.

Before	110	120	125	132	125
After	120	118	125	136	121



Solution: Let μ be the mean of population of differences.

- **Null Hypothesis** $H_0: \mu_1 = \mu_2$ i.e, no change in B.P.
- **Alternative Hypothesis** $H_A: \mu_1 \neq \mu_2$ i.e, no change in B.P.
- **Level of Significance:** $\alpha = 0.01$
- **Computation :** Differences d_i 's (before and after drug) are
-10, 2, 0, 14, 4

$$\begin{aligned}\bar{d} &= \frac{-10 + 2 + 0 + -4 + 4}{5} \\ &= \frac{-8}{5} = -1.6\end{aligned}$$

$$\begin{aligned}S^2 &= \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 \\ &= \frac{1}{4} \sum_{i=1}^5 (d_i - \bar{d})^2 \\ &= \frac{1}{4} [(-10 + 1.6)^2 + (2 + 1.6)^2 + (0 + 1.6)^2 + (-4 + 1.6)^2 + (4 + 1.6)^2] \\ &= \frac{123.20}{4} = 30.8 \\ S &= \sqrt{30.8} = 5.55\end{aligned}$$

- **Test statistic:** The test statistic is t which is calculated as

$$\begin{aligned}t &= \frac{\bar{d} - \mu}{S / \sqrt{n}} \\ &= \frac{-1.16}{5.55 / \sqrt{5}} = -0.645\end{aligned}$$

Calculated $|t|$ value is 0.645

Tabulates $t_{0.01}$ with $5-1 = 4$ degrees of freedom is 3.747.

Since calculated $t < t_{0.01}$, we accept the Null hypothesis and conclude that there is no significant change in blood pressure.

R code:-

```
> x <- c(110,120,125,132,125)
> y <- c(120,118,125,136,121)
> t.test(x,y,paired=TRUE)
```

Paired t-test

```
data: x and y
t = -0.64466, df = 4,
p-value = 0.5543
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.490956  5.290956
sample estimates:
mean of the differences
 -1.6
```

T-test for difference of two population means :-

With a two-sample t-test, we compare the population means to each other and again look at the difference. We expect that $\bar{x} - \bar{y}$ would be close to $\mu_1 - \mu_2$. The test statistic will use both sample means, sample standard deviations, and sample sizes for the test.

A two-sample t-test follows

- Write the null and alternative hypotheses.

- State the level of significance and find the critical value. The critical value, from the student's t-distribution, has the lesser of n_1-1 and n_2-1 degrees of freedom.
- Compute the test statistic.
- Compare the test statistic to the critical value and state a conclusion.

$$t = \frac{\bar{x} - \bar{y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \quad \text{or} \quad S^2 = \frac{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}{n_1 + n_2 - 2}$$

Problem:- Two horses A and B were tested according to the time (in seconds) to run a particular track with the following results.

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether the two horses have the same running capacity.

Solution:- Given $n_1=7$ and $n_2 = 6$

We first compute the same means and standard deviations.

\bar{x} = Mean of the first sample

$$= \frac{1}{7} (28 + 30 + 32 + 33 + 33 + 29 + 34) = \frac{1}{7} (219) = 31.286$$

\bar{y} = Mean of the second sample

$$= \frac{1}{6} (29 + 30 + 30 + 24 + 27 + 29) = \frac{1}{6} (169) = 28.16$$

x	$x - \bar{x}$	$(x - \bar{x})^2$	y	$y - \bar{y}$	$(y - \bar{y})^2$
28	-3.286	10.8	29	0.84	0.7056
30	-1.286	1.6538	30	1.84	3.3856
32	0.714	0.51	30	1.84	3.3856
33	1.714	2.94	24	-4.16	17.3056
33	1.714	2.94	27	-1.16	1.3456
29	-2.286	5.226	29	0.84	0.7056
34	2.714	7.366			
219		31.4359	169		26.8336

$$\begin{aligned} \text{Now, } S^2 &= \frac{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}{n_1 + n_2 - 2} \\ &= \frac{(31.4358 + 26.8336)}{7 + 6 - 2} = 5.23 \end{aligned}$$

$$\text{Therefore } S = \sqrt{5.23} = 2.3$$

- Null Hypothesis $H_0: \mu_1 = \mu_2$
- Alternative Hypothesis $H_A: \mu_1 \neq \mu_2$
- Level of Significance: $\alpha = 0.05$



- **Computation:** $t = \frac{\bar{x} - \bar{y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{31.286 - 28.16}{(2.3) \sqrt{\frac{1}{7} + \frac{1}{6}}} = 2.443$

Tabulates $t_{0.05}$ with $7+6-2 = 11$ degrees of freedom at 5% level of significance is 2.2

Since calculated $t > t_{0.05}$, we reject the Null hypothesis and conclude that there is no significant change in blood pressure.

ANOVA:- (ANALYSIS OF VARIANCE)

When we have only two samples we can use the t-test to compare the means of the samples but it might become unreliable in case of more than two samples. If we only compare two means, then the t-test (independent samples) will give the same results as the ANOVA. Anova is performed with F-test.

Null hypothesis H_0 : There are no differences among the mean values of the groups being compared (i.e., the group means are all equal)-

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

Alternative hypothesis H_1 : (Conclusion if H_0 rejected)?

Not all group means are equal (i.e., at least one group mean is different from the rest).



ANOVA one-way classification:-

Step 1: Total number of all observations

$$T = \sum_i \sum_j X_{ij}$$

Step 2: Correlation factor

$$cf = \frac{T^2}{N} = \frac{T^2}{r \times s}$$

Step 3: Total sum of squares

$$TSS = S^2T = \sum_i \sum_j X_{ij}^2 - cf$$

Step 4: Treatment sum of squares

$$TrSS = S^2Tr = \sum \frac{T_j^2}{N} - cf$$

Step 5: Error sum of squares

$$ESS = S^2E = TSS - TrSS$$

Source of variable	d.f	Sum of Squares	TSS	F-Test
Treatment (between sample)	k-1	$S^2Tr = \sum \frac{T_j^2}{N} - cf$	$S^2Tr = \frac{ST_r^2}{k-1}$	$F_{cal} = \frac{S^2Tr}{S^2E}$
Error	n-k	$S^2E = TSS - TrSS$	$S^2E = \frac{S^2E}{n-k}$	