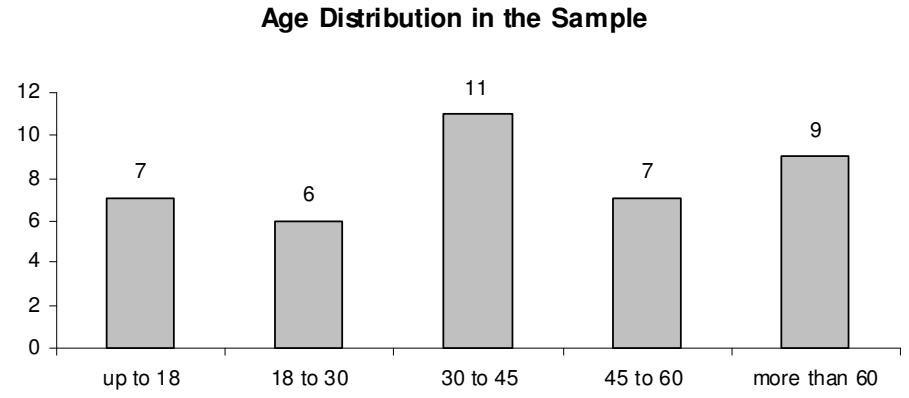
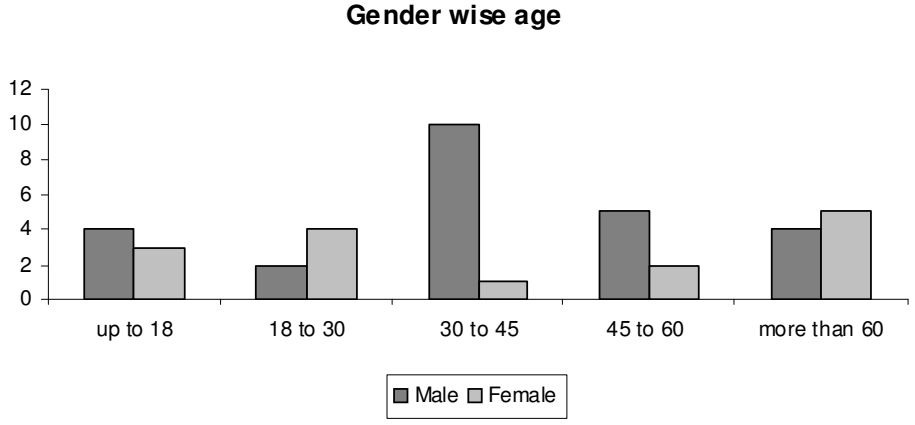
9.3 DATA PRESENTATION

Besides tables, data can also be represented by suitable diagrams. Most widely used diagrams are bar diagrams, histograms and pie diagrams. A bar diagram or bar graph is a chart with rectangular bars with lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. The frequency distribution of age classes can be represented by bar diagrams as shown below



A bar diagram is very useful if you are trying to record certain information whether it is continuous or not continuous data. It is nevertheless more suitable for representing discontinuous data. There can be variations in the bar diagrams as well. For instance if one category has more than one value, for instance the male and female as shown in the cross table above, we use clustered bar diagrams. The appropriate bar diagram for the cross table given above will be as shown below





CHECK YOUR PROGRESS

Q.1: What is a bar diagram?

.....

.....

Q.2 : What do you mean by pie chart?

.....

.....

Q.3 : Describe histogram.

.....

.....

9.4 HYPOTHESIS TESTING, TESTS & INTERPRETATIONS

Research starts with an initial statement known as **hypothesis**. In statistics, the initial hypothesis is called **null hypothesis**. The hypothesis or null hypothesis to be precise is validated through data and after validation, the hypothesis turns into a **theory** which is nothing but a *broad generalisation* for entire population. Clearly, therefore, statistical inference is a key element in theory building. In this section we will briefly discuss about important elements of the statistical inference.

Let us start with the clear articulation of our exact problem with the aid of an easily comprehensible example. Suppose that we have a class of students. The normal assumption of deciding about promoting a student to a higher class requires that he or she acquires knowledge adequately at that level. For knowing whether the student has acquired the required knowledge adequately, a test is held. Based on the marks obtained in the test we decide about his or her promotion to a higher class. Note that this essentially resembles to the process of statistical testing. The process can be illustrated as under:

The **null hypothesis** here is that “the student is to not be promoted”.

The type I error is also known as **false positive** and the type II error is known as **false negative**. Given this two types of errors, one can also find their rates of occurrence. For instance, say out 100 how many times they occur. If any one them occurs 5 times, then, obviously the rate is given by 5 percent. The rate of type I error i.e. the rate of false positive is known as **significance or level of significance** (also denoted by á) while the rate of type II error i.e. the rate of false negative is known as **power** (denoted by â). Ideally, we want low levels of these two rates for safe generalisation. However, these two has a trade-off one can be lowered at the expense of the other.

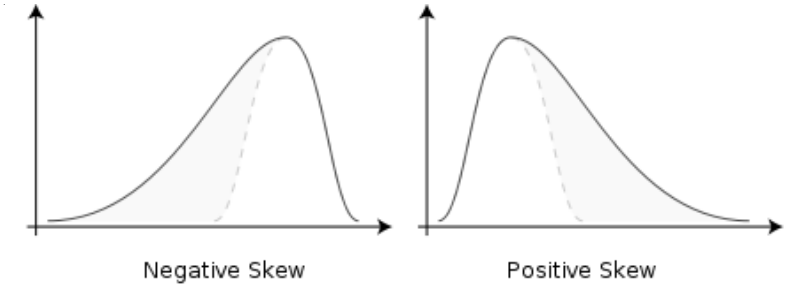
Further note that other way of expressing the level of significance is the **level of confidence**. If level of significance is say 5 percent the rate of type one error is 5 out of every 100. Speaking other wise we are correct in 95 out of 100, indicating thereby our level of confidence. Hence, 1-level of significance is the level of confidence.

For all statistical tests there is a predefined probability distribution of critical values for different levels of significances. These critical values provide us minimum evidence required to accept or reject the null hypothesis as the case may be. One has to compare one's own observed test value against the theoretical critical value for possible acceptance and rejection. The associated levels of significance will tell us the probability of making type I error while making decisions based on such results.

Recall our student and examination example. As there are different examinations for different occasion viz. annual examination for regular promotion, entrance examinations for getting admissions etc., there are different statistical tests as well suited for different occasions and situations. Below we will discuss very briefly some of the most popular and useful tests.

Most of the statistical analysis assumes **normality** on the part of the distributions under the study. Simply speaking, normality refers to an average situation of any phenomenon where "extremes are rare and average is more common". This means that in any situation, too high or too

maximum as always, the mean moves towards the tails due to the impact of extreme value. The median stays in between the mode and the mean. Empirically, in such distributions the distance between the mean and median is about one-third of the distance between the mean and mode. Karl Pearson expressed this relationship as: Mode = Mean - 3 [Mean - Median]. This relationship between mean, median and mode with skewness is very important in analysing the data and making statistical interpretations.



As has been pointed out the normal curve is symmetrical about the mean. Apart from the symmetry, there is another very useful property of the normal curve. The area under the normal curve representing probabilities of the values expressed in terms of the mean and standard deviations are standardised and already tabulated. Any value, say *x*, of a distribution can be expressed in terms of the mean and standard deviation of the distribution as

$$Z = \frac{x - \bar{x}}{\sigma}$$

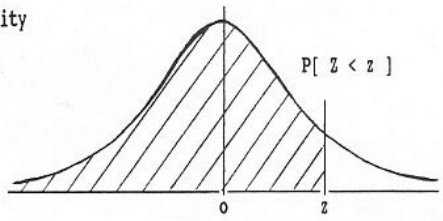
Where $\bar{x}$ and σ are the mean and standard deviation of the distribution. Recall that this simply indicates change of origin and scale. In this transformation we are re-scaling the values taking the mean as the origin and standard deviation as the interval. This scale is given by the horizontal axis of the normal curve where the mean can be treated as the origin and intervals are given by the standard deviations. The values of *Z* actually put the values of *x* in this new scale implying *how many standard deviations away is the value from mean* in the axis. For example, if mean is 5 and standard deviation is 3, then *Z* transform of 20 will be (20-5)/3= 5, which simply means that original value of 20 when put on the *Z* scale will lie 5 standard deviations away from the mean to its right side. This Z value is called **standard normal variable.**

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e. $P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964

Working Example 1

Suppose that legal speed limit of vehicle in a city is 40 km. A study was conducted to find out how many vehicles actually violate the legal norm. The study was based on a random sample of 200 vehicles. It was found that mean speed limit was 50 and standard deviation was 5. Also distribution of the sample speed is found to be normal. What can be concluded on the basis of the information?

Q.5 : What is test value?

.....

.....

Q.6 : What is level of significance?

.....

.....

Q.7 : What are positive skew and negative skew?

.....

.....

9.4.1 t-Test and its Uses

In quantitative research we very commonly compare means. For instance, we may be interested in studying income differences amongst working male and females. Or we may be interested in studying IQ levels of boys and girls etc. Here we, typically, want to compare *means* i.e. mean male income and mean female income or mean boy IQ and mean girl IQ etc. Obviously, since we are studying on sample basis there will be some differences. However, there is a limit to that. If differences are large enough which is not due to sampling and all then the difference is called **statistically significant**. When test result is found to be significant (at some level of significances) we simply mean that difference is genuine and there must be some reason to it. **Student t test** or popularly known as **t-test** is normally used to find out statistically significant differences amongst **two means**.

The t-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name). He published the test in Biometrika in 1908, but was forced to use a pen name by his employer, who regarded the fact that they were using statistics as a trade secret.

The t-test can be applied to various situations. Primarily, three situations emerge in research yielding three different types of t-tests. Imagine a situation we have conducted a sample survey

Further, there are other two issues involved. The two samples can be of same size or different size. Also the two distributions can have equal variances or unequal variances. Depending on these factors actual formula of t-statistics can vary. These days, the testing is done on computers which handle these complexities quite easily.

Any case, one needs to understand the important procedural steps in conducting t-test. The test is applied to *compare two means* with the null hypothesis that there is no difference in them. When the test result is found significant we reject the null and say the there is significance difference in means compared. Then one must try to attribute reasons as what might have caused the difference.

9.4.2 F-Test and its Application

As has been pointed earlier the t-test can be applied only for comparing two means. There can be situations when researchers may have to compare more than two means. Then F-test becomes handy. The name was coined by George W. Snedecor, in honour of Sir Ronald A. Fisher. Fisher initially developed the F-statistic as the variance ratio in the 1920s. It is, therefore, also applied to test the fit of model to the data, in explaining proportion of variance explained by a model out of the total variance observed.

It is important to observe that all the three tests – Z, t and F hold good only under certain assumptions. Most important among them is the assumption of normality. Unless distribution is normal these tests, although widely used, can not be applied to data. Two issues, therefore, merit some discussion: how do we know that our data confirm to normality and second if this is not so what is to be done.

First of all, let us confess that in real life situations, we hardly get data which are perfectly normal. In other words, in real life we are bound to have some degree of deviations from normality. Therefore, one has to see whether deviation is significantly large enough to cause problems in analyses. There are two ways which can detect deviations from normality viz. graphical and formal test of normality. Most widely used graphical method is the **box-plot**

been 14 per hour where as we have observed the actual as given by the series. Evidently, there is a deviation from fairness, which can be given as below:

1 **Hour**	*2* **O**	*3* **E**	*4* **(O-E)**	*5* **$(O-E)^2$**	*6* **$(O-E)^2/E$**
1	12	14	-2	4	0.29
2	13	14	-1	1	0.07
3	16	14	2	4	0.29
4	13	14	-1	1	0.07
5	20	14	6	36	2.57
6	17	14	3	9	0.64
7	9	14	-5	25	1.79
8	12	14	-2	4	0.29
Total	**112**	**112**	**0**	**84**	**0.75**

The column 4 gives the deviations from fairness. Note that total deviation is 0. Recall that this is because of the property that sum of deviations from the mean is 0. To do away with the problem, we square the deviations and express that as proportion to the expected which is given in the column 6. The total 0.75 is known as the value of chi-square and it is formally given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Like any other statistical test, the critical values for this test are also tabulated. The tabulated value for *n-1* degree for freedom i.e. 7 in this case at 5 percent level of significance is, for instance, 12.47. Since in our case the calculated value is found far less than what is tabulated, we can safely take the distribution as fair and hence there is no reason to believe that number of accidents is bias towards 5th hour. The high value observed in this distribution at 5th hour is only due to chance and not because of any external reason.

The second use of the chi-square lies in interpreting contingency tables. A contingency table classifies the distributions according to sub-categories. Simplest type of contingency table takes a 2 row 2 column form as given below:



CHECK YOUR PROGRESS

Q. 8: What is independent t-test?

.....

.....

Q. 9: What is box plot method?

.....

.....

Q. 10: What do you understand by chi-square method?

.....

.....



9.5 LET US SUM UP

- In this chapter we have discussed about various ways of diagrammatically representing data. We also have introduced the processes and principles of testing of hypothesis. We have introduced various tests like z test, t test, F test and chi-square test. We have also discussed their utility and applications. Further the chapter also discusses how data can be interpreted for reporting using the various test procedures.
- Bar charts also look a lot like a **histogram**. They are often mistaken for each other. A histogram is a graphical representation of data showing a visual impression of the *distribution* of data. Research starts with an initial statement known as **hypothesis**. In statistics, the initial hypothesis is called **null hypothesis**. The statistical analyses which require the normality assumptions to be fulfilled are known as **parametric statistics**. There are, of course, some analyses which do not need the assumption of normality, and they are called **non-parametric statistics**. The hypothesis or null hypothesis to be precise is validated through data and after validation, the hypothesis turns into a **theory** which is

Ans to Q No 7 : If, however, there is a deviation from normality mean, median and mode values will differ and it will produce longer tails towards either side. When the tail will be elongated towards positive side we call it **positive skew** and when the tail elongates towards negative side we call it **negative skew**.

Ans to Q No 8 : When t-test is applied to two independent samples it is called independent t test.

Ans to Q No 9 : The box-plot is a quick way of examining one or more sets of data graphically.

Ans to Q No 10 : A chi-squared test is a statistical hypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true.



9.8 POSSIBLE QUESTIONS

Q.1: Draw a bar diagram for the following table.

Income Groups	Rural (%)	Urban (%)
Less than 10000	54	42
10000 to 40000	26	39
40000 to 80000	12	10
More than 80000	8	9

Q.2: Draw a pie diagram for the following data.

Social Groups	Population (%)
General	62
Scheduled Caste	13
Scheduled Tribe	19
OBC	6