10/ Medical Statistics

Sometimes we want to explore new information about facts present in the universe. Any information regardless of its motive depends partly on real observations and partly on reasoning. The inference drawn either from observation or reasoning or from both is named as **hypothesis.** This hypothesis or *inference* in any case cannot be treated as an absolute result or value. In a large population it is not possible to observe all the items. So statistical methods are developed to estimate the *error* or *deflection* from the true or real value when we are taking inference from a sample for a population.

The statistical methods are utilised in interpreting the results which are at the mercy of numerous influences and obtained from different experimental procedures. The object of the statistical analysis is to isolate and measure the effects of individual influences. To get a correct statistical analysis we have to see that we are comparing groups like with like that means we have not overlooked any relevant factor which is not present in anyone of the two groups.

Experiment with biological material is much more complex. Most of the biological parameters are interrelated and influenced by many other factors. When we measure a parameter, the readings are scattered over a wide range, even when we think that the affecting conditions are constant. Again when the readings are taken after altering some of the affecting conditions, the range of the readings often overlap the former instead of being a completely separate one. Under such circumstances, statistical assessment can help in several ways:

1. To choose a value which is representative of each group of readings.
2. To describe what is the nature of the variability within the group.
3. To decide whether a difference between two groups is a real one or it could have occurred by chance.

I. Probability

When we are conducting some experimental observations, first of all we have to consider that whether the result is obtained only due to chance or coincidence, or it is appearing as the real or true one. To meet this query, we have to find out the natural probability. First, say, where n is the number of all possible results of some action, which are of equal possibility to occur and r is the number of results which will satisfy some particular necessity or condition, then the probability of satisfaction of the requirement will be (r/n). As for example, a die can fall with any of its six sides ($n=6$) keeping uppermost. If the necessity is such that the side marked with 4 should come uppermost then only one of the six sides will satisfy the requirement, i.e., $r=1$. So in this case the probability is (1/6). When we consider r and n for an ideal die then any of the numbers from 1 to 6 is equally likely to come upwards. In that case the P is equal to (1/6). This type of preassumed probability is named as a *priori probability.* But the probability may also be obtained by actual throwing of the die for a number of times (more the number of throwing better or more nearer will be the P value to the ideal one). The probability obtained with this process is named as a *posteriori probability*. In practice, this is a more complex process than the priori probability. So in the case of a die a priori probability is (1/6), i.e., (r/n) but a posteriori probability is (r_x/n_x) which may not be exactly equal to (1/6).

Probability of more than one event. The probability of occurring multiple events at a time is same as above, i.e., in case of a single event, (r/n), when n is the possible way in which all the combinations of events can occur and r the number of combinations that fulfils the requirement. These multiple events may be dependent or independent of each other.

Probability of mutually dependent events. When the requirement is so that any of the mutually exclusive results (events) satisfy then the individual probabilities are to be added together to get the final probability. In other words, when individual probabilities are (1/6), (1/6), (1/6) for appearing upwards anyone of the numbers more than three of a die, i.e., 4, 5 and 6 respectively; then probability of appearing anyone number more than three in a die is $(1/6)+(1/6)+(1/6)=(3/6)$. In summary, this can be stated as—if either a or b or c, add the probabilities of a, b and c.

Probability of independent events. If the requirement is so that the concurrence of two or more events, each of which can occur independent of the other then the probability of concurrent events can be obtained by multiplying the individual probabilities. Say for example, when one coin and one die are tossed together and the requirement is that the head of the coin and the side of the die containing number 4 should come upward; then individual probabilities are (1/2) and (1/6) respectively; and the probability of concurrence of these events is $(1/2)\times(1/6)=(1/12)$. In short this may be stated as—if only a as well as b, as well as c will satisfy the requirement, then multiply the probabilities of a, b and c.

Calculation of n and r to find out the probability of larger systems. In larger complicated systems it is more convenient to calculate n and r by the method or formula of permutation and combination.

All possible different arrangements which can be made out of n things by taking some (m) or all (n) of them at a time is called their **permutation** (${}^nP_m$ or ${}^nP_n$) respectively, which is stated as

$$ {}^nP_m = n(n-1)\ (n-2) \cdots (n-m+1) $$

$$ {}^nP_m = {}^nP_n = {}^n = n! = n(n-1)\ (n-2) \cdots 3.2.1, \text{ when } n=m. $$

Now again, when each combination of m objects from n objects differs only of content and not of inner arrangement, it is called **combination** (${}^nC_m$).

$$ {}^nC_m = \frac{{}^nP_m}{m!} = \frac{n!}{m!\ (n-m)!}. $$

For example, the number of ways by which any four cards can be drawn from a packet of playing cards is ${}^{52}P_4 = 52\times51\times50\times49 =$ 64,97,400, when the inner arrangement or order is considered, ${}^{52}C_4 = \frac{{}^{52}P_4}{4!} = \frac{52\times51\times50\times49}{4\times3\times2\times1} = 2{,}70{,}725$, when the inner arrangement or order is not considered.

So the probabilities of getting 4 jacks together, with or without considering the inner order are (1/64,97,400) or (1/2,70,725) respectively (as here $r=1$ in both the cases and $n=64{,}97{,}400$ and $2{,}70{,}725$ respectively).

Now in practice, more the number of repetition of the experiment the P value will be more nearer to the theoretical or ideal one. Smaller the number of repetition larger will be the discrepancies which can also be calculated with the following formula:

$$P=\frac{n!}{r!\,(n-r)!}\cdot p^{r}\cdot(1-p)^{n-r}$$

when p is the probability of a positive result in some situation, the probability, P of getting r positive results in n trials.

For example, in case of a coin, the probability of getting four times head upwards in 8 trials will be as follows:

$$P=\frac{8!}{4!\,(8-4)!}\cdot(\tfrac{1}{2})^{4}\cdot(1-\tfrac{1}{2})^{8-4}=\frac{8!}{4!4!}\cdot(\tfrac{1}{2})^{4}\cdot(\tfrac{1}{2})^{4}=\frac{8!}{4!4!}\cdot(\tfrac{1}{2})^{8}=(35/128).$$

Much of statistics involves the use of test to work out what is the probability of a particular difference in events occurring by chance. Then with an arbitrary definition it can be stated that the result is whether or not **statistically significant.** When $P \leqslant 0{\cdot}05$, i.e., one in twenty trials, is significant and $P \leqslant 0{\cdot}01$, i.e., one in hundred trials, is taken as highly significant. Still $P \leqslant 0{\cdot}01$ does not mean that there is a real difference between two sets of results but only that such a difference can occur one in hundred trials by chance.

II. Frequency distribution

The experimental observations may be divided mainly into two groups, **quantitative,** dealing with some property or change of property in quantity and **qualitative,** some sort of property (quality). Generally the number of observations is smaller when dealing with qualitative observations, and very large when dealing with quantitative observations. When experimental results are obtained, they appear as mere collection of data. In many cases it is therefore advisable to prepare smaller groups of them by dividing the range of data into smaller subdivisions according to some standards and then to obtain a frequency distribution. So frequency distribution of a set of data is the frequency of occurring of the data in different smaller subdivisions of the range of data.

Preparation of a frequency distribution table (Fig. 10.1). The frequency distribution table can be prepared by jotting down all the numerical values in order of increasing size which are represented in the series of measurements in the first column, next the mean value of each range. In the *occurrence* column one vertical stroke is marked for each occurrence of the value against the respective range of subgroup and a group of five is made with four such vertical strokes and a diagonal one across the first four (Fig. 10.1). As for

Actual range of groups	*Mean value*	*Occurrences*	*Total*
1·48 - 1·50	1·49	|	1
1·50 - 1·52	1·51	||||	4
1·52 - 1·54	1·53	𝍸 ||	7
1·54 - 1·56	1·55	𝍸 ||||	9
1·56 - 1·58	1·57	𝍸 |	6
1·58 - 1·60	1·59	|||	3

Total number of data is 30.

Fig. 10.1. Preparation of frequency distribution chart.

example, the results obtained by measuring the height (in metres) of 30 human adults of age group 20–25 years: 1·51, 1·54, 1·55, 1·59, 1·48, 1·50, 1·55, 1·53, 1·50, 1·55, 1·55, 1·52, 1·53, 1·54, 1·53, 1·56, 1·52, 1·57, 1·53, 1·57, 1·52, 1·54, 1·51, 1·54, 1·55, 1·56, 1·57, 1·56, 1·59, 1·59.

The frequency distribution table is prepared with these results by dividing into smaller ranges as 1·48 to 1·50; 1·50 to 1·52, etc., along with a mean value in each case and the frequency of occurring the results in different subgroups are found.

Frequency polygon. The frequency polygon (FIG. 10.2). is nothing but the graphic representation of the frequency distribution table. The graph is to be drawn with the mean values of the range of subgroups along the abscissa and the frequency distribution along the ordinate. If we increase the number of results and divide into smaller subgroups then the polygon will be more accurate and smoother one which is then named as **frequency distribution curve.** When the frequency is equally distributed on the both sides of the midpoint (highest frequency) the curve will be bell-shaped and named as **normal, Gaussian or symmetrical distribution curve** (FIG. 10.3 A). If there are only two possible alternatives for a parameter (e.g., head and tail of a coin; Rh^+ and Rh^-) and we can precisely state the number of times one or the other occurs. In such a case the distribution is named as **binomial** which may be symmetrical or asymmetrical. When the slope of the curve is less at the positive side (right side)—it is **asymmetrical, positively skewed** (FIG. 10.3 B), the reverse one is the asymmetrical, **negatively skewed** (FIG. 10.3 C) distribution. When the frequency

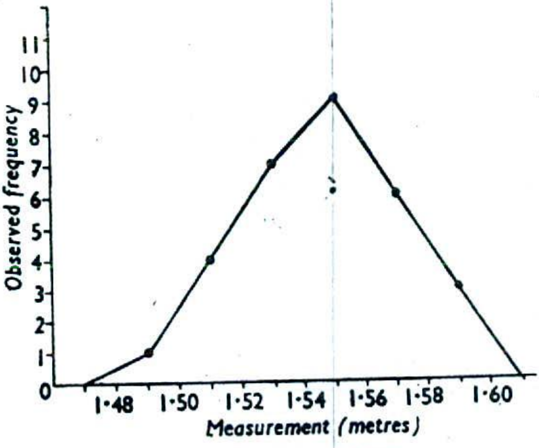
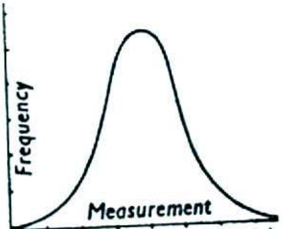
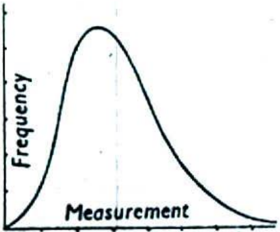


FIG. 10.2. Frequency polygon.



Normal or Gaussian symmetrical distribution (Bell-shaped)

FIG. 10.3A.

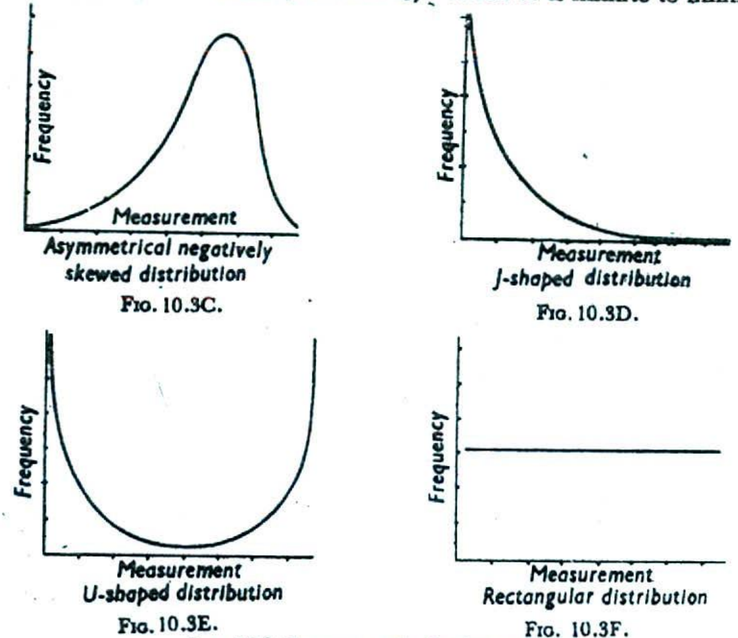


Asymmetrical positively skewed distribution

FIG. 10.3B.

FIG. 10.3. Frequency distribution curves (contd.).

polygon is tending to approach a J-shape, i.e., frequency tends from infinity to zero but remains finite as measurement increases, it is **J-shaped** or **Poissonian distribution** (Fig. 10.3 D). When it is infinite to infinite—



Asymmetrical negatively skewed distribution

Fig. 10.3C.

J-shaped distribution

Fig. 10.3D.

U-shaped distribution

Fig. 10.3E.

Rectangular distribution

Fig. 10.3F.

Fig. 10.3. Frequency distribution curves.

U-shaped (Fig. 10.3 E), i.e., at both the highest and lowest values of the parameter the frequency is infinite. When the frequency is equal throughout the groups of data, then the curve will be parallel to the horizontal axis—**rectangular** (Fig. 10.3 F).

Frequency histogram. The frequency with which each of the values in the series occurs is therefore qualitative statistics, in other words, each refers to the number of individual measurements falling into a class having well-defined limits. The graph can be done by putting the ranges in place of the measurements, e.g., 1·51 metres = 1·50–1·52 metres and forming a rectangle with the frequency, i.e., the rectangle is formed with the range as the

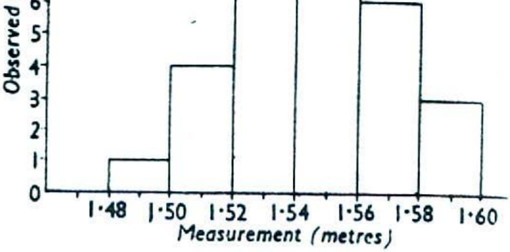


Fig. 10.4. Frequency histogram.

base and the frequency distribution as the height. This type of graphic representation is named as **frequency histogram** (FIG. 10.4).

The probability of occurring any individual measurement that lies within any specific limits, can be inferred from the frequency polygon or histogram drawn from a set of data. The probability is proportional to the frequency for each of the stated values representing the midpoint of the range in case of the frequency polygon. But in case of the histogram the probability is proportional to the rectangle made of the range of values and frequency observed. Now normally the ranges of values are equal to each other and then, the probability is proportional to the frequency observed.

III. **Averages** (FIG. 10.5)

1. **Arithmetic mean.** When several readings are obtained from the same parameter (object) or of the similar kind of objects, then usually an average (mean) is taken from all the readings. One of such averages which is most commonly used is the arithmetic mean. Generally the term 'average' or 'mean' is used for the arithmetic mean. Now for example when the set of data is x_1, x_2, x_3, $\cdots$, x_n; n being the number of data (readings); $\Sigma(x)$ stands for summation of the whole set of data (i.e., $x_1+x_2+x_3+\cdots+x_n$); then the arithmetic mean ($\bar{x}$) can be obtained with the following formula:

$$\bar{x}=\frac{\Sigma(x)}{n}.$$

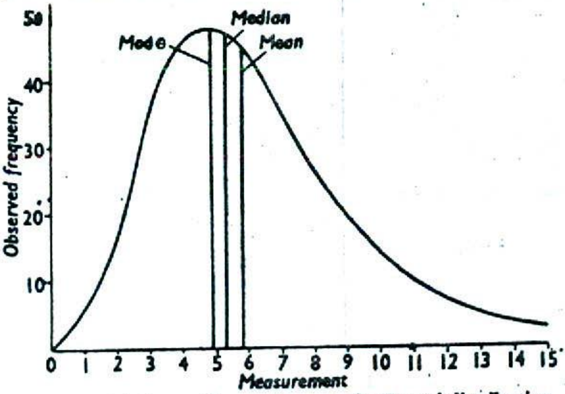


FIG. 10.5. Mode, median and mean of a skewed distribution.

2. **Mode.** There are some conditions (requirements) where the arithmetic mean does not serve the purpose. Say for example, one businessman wants to stock ready-to-wear garments of standard size which will fit maximum number of customers. In this case he will be interested of the size of customers most frequently occurring in the population and notin the arithmetic mean. So the mode is the most frequently occurring value of x; in other words, the mode is the highest point of the frequency polygon.

3. **Median.** It is the third kind of average which divides a set of data into two equal halves, one containing the values lower than this value and the other containing the higher values. The formula to find out the position of the median is $\{(n+1)/2\}$, in other words, $\{(n+1)/2\}$th value of x. For example, there are 107 items in a set of data, then $\{(107+1)/2\}$=54th value in the range of data is the median.

If the distribution of the frequency is a symmetrical one then the median coincides with both the mean and the mode. If it is a positively or negatively skewed one then the median will lie in between the mode and the mean. The mode is the peak of the curve and the mean towards the long tail end.

4. **Geometric mean.** This kind of average is used when dealing with ratios, indices or any such other relative numbers.

$$M\bar{g} = {}^{n}\sqrt{x_1 . x_2 . x_3 . \cdots . x_n}.$$

5. **Harmonic mean.** This is used when dealing with rates and speeds.

$$M\bar{h} = \frac{n}{\Sigma(1/x)}.$$

In summary, an average can have a variety of proportions, but it is not always possible to show all of them in one value. It may be the one nearest to all experimental values (arithmetic mean), or the value that divides the experimental values into two equal halves of larger and smaller values (median), or the value that occurs most frequently (mode).

6. **Average of averages.** Sometimes it needs to have a grand average of the averages obtained from smaller groups or samples. To get the grand mean the size of each sample must be considered when the sample size is not constant for all of them. Say for example, there are r sets of samples, having means as m_1, m_2, m_3, ..., m_r with n_1, n_2, n_3, ..., n_r number of values respectively. Then $\{$as $S(x) = m.n\}$ the sum of the results of the first sample is $m_1 . n_1$; the sum of all the results of all the samples is

$$S(mn) = m_1n_1 + m_2n_2 + m_3n_3 + \cdots + m_rn_r.$$

Total number of results or data of all the samples together is

$$S(n) = n_1 + n_2 + n_3 + \cdots + n_r.$$

The grand arithmetic mean (M) is then

$$M = \frac{S(mn)}{S(n)} = \frac{m_1n_1 + m_2n_2 + m_3n_3 + \cdots + m_rn_r}{n_1 + n_2 + n_3 + \cdots + n_r}.$$

If the size of the samples is same for all the cases (i.e., $n_1 = n_2 = n_3 = \cdots = n_r$) then

$$M = \frac{n(m_1 + m_2 + m_3 + \cdots + m_r)}{rn} = \frac{S(m)}{r}.$$

7. **Averages as ideal values.** It is already stated that the larger the number of individual values, the more likely that the mean will attain the ideal value as the dispersed values will balance each other. The averages obtained from the small group of estimates may not be very near to the true mean, so they are denoted by the Roman letters as m_1, m_2, etc.; to differentiate from the ideal value which is denoted by the Greek letter μ.

IV. Deviations or scatters

Suppose we are measuring the systolic blood pressure of a group of male individuals by means of two methods. With the first method the lowest reading (datum) is 96 mm of Hg and highest is 104 mm of Hg; and with the second process lowest is 45 mm of Hg and highest is 155 mm of Hg. Both the groups are normally distributed with an average 100 mm of Hg. Now what method is to be recommended where only one sample or reading is possible. To answer this question the following statistical analysis may help us. It is seen that the individual readings in case of the first method are more nearer to the true mean than that of the second one. In that case with the second method, a single reading may give a impression of the situation with a larger error. These differences or distances from the mean are named as *deviations* or *scatters*.

1. **Range of data.** The simplest and commonest measure of scatter is the range, which is the interval between the lowest and the highest value of a set of data, e.g., in the above paragraph the range of data is 96–104 with the first method and 45 to 155 in the second one. The end points have got no definite destination. the size of the range to some extent depends on the size of the sample, i.e., a larger sample is more likely to contain more extreme values making the range larger than a smaller sample. So it is not at all a reliable measure of scatter

2. **Mean deviation (M.D.).** It is a better way of measuring the scatter as it takes account of all the numbers of the sample and its value does not depend on the size of the sample, but on the size of the deviations respectively. The difference between any individual reading and the arithmetic mean in a set of data is named as the *deviation*. The sum of the deviations without considering the sign (of the difference) divided by the number of values is called the *mean deviation*.

$$\text{M.D.} = \frac{\Sigma\,|\,x-\bar{x}\,|}{n}$$

when $|x-\bar{x}|$ = the difference ignoring the sign,

and n = number of values.

3. **Variance and invariance.** The sum of the squared deviations will not be affected by the sign of the individual differences as they are squared. The variance is the sum of the squares of the individual deviations (differences) divided by the number of the values of a set of data,

$$\text{V (variance)} = \frac{\Sigma(x-\bar{x})^2}{n}.$$

It is largest in a population in which the individuals vary a lot.

The invariance is the reciprocal of the variance which is largest in a population where the individual readings vary minimum,

$$\text{i.e., I (invariance)} = \frac{n}{\Sigma(x-\bar{x})^2}.$$

4. **Standard deviation.** The standard deviation is the square root of the variance, i.e., standard deviation $(\sigma) = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$. As the ideal or true mean of a population can be obtained only approximately from a sample so the true standard deviation can be estimated also approximately from the available sample. With this appreciation the standard deviation can best be estimated with the following formula $\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{(n-1)}}$. It is the

most commonly used expression of deviations. It has the advantages as that of the mean deviation, and some other advantages specially when dealing with the samples of a normally distributed population. With such distribution, one standard deviation (σ) at both positive and negative side of the mean ($\bar{x}\pm\sigma$) will cover about two-thirds or 66% of the total number of values, similarly ($\bar{x}\pm2\sigma$) will cover 95% and ($\bar{x}\pm3\sigma$) will cover 99% in approximate of the total number of values (Fig. 10.6). The arithmetic labour that is involving in calculating the standard deviation can be minimised with the following systematic tabular form:

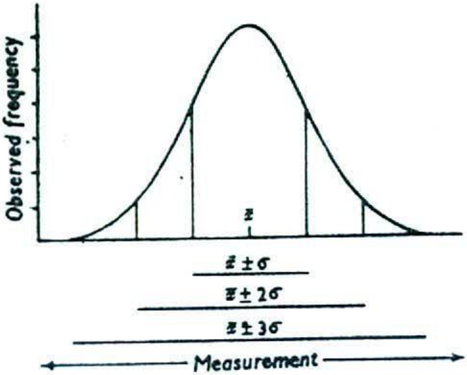


Fig. 10.6. Relationship between the standard deviation and the normal or Gaussian frequency distribution.

No. of observations	Individual observations (x)	Deviation ($x-\bar{x}$)	Squared deviations $(x-\bar{x})^2$	Calculation of S.D. (σ) [σ = *sigma* (small)]
1	8	−2	4	$\Sigma(x-\bar{x})^2=60$
2	10	0	0	
3	9	−1	1	$\sigma=\sqrt{\frac{\Sigma(x-\bar{x})^2}{(n-1)}}$
4	11	1	1	
5	12	2	4	
6	10	0	0	
7	7	−3	9	$=\sqrt{(60/9)}$
8	6	−4	16	
9	13	3	9	$=\sqrt{6{\cdot}66}$
10	14	4	16	$=\pm2{\cdot}58$
Total	$\Sigma(x)=100$ $\bar{x}=\frac{100}{10}=10$	$\Sigma(x-\bar{x})$ $=0{\cdot}0$	$\Sigma(x-\bar{x})^2=60$	

The above-mentioned method for calculating the standard deviation is suitable for smaller groups of results but not for large-sized samples. For large samples the following table is more appropriate:

Mean value of the smaller ranges in cm	Frequencies	Deviations	Squared deviations	Product of respective deviation and frequency	Product of squared deviation and frequebcy	Working
149	4	—5	25	—20	100	Working mean=154
151	9	—3	9	—27	81	Mean=154 +0·12=154·12
153	12	—1	1	—12	12	Correction for mean $(6^2/50)=0\cdot72$
155	14	1	1	14	14	Corrected sum of squared deviations...
157	8	3	9	24	72	354—0·72=353·28 sq. cm.
159	3	5	25	15	75	$\therefore$ S.D.$=\sqrt{\dfrac{353\cdot28}{49}}$
Total	50	—	70	—6	354	$=\pm2\cdot68$ cm

V. Sampling

A true mean is the mean of a infinitely large number of readings of a parameter. In experiments it is not possible for us to collect an infinite number of readings of a parameter, so we have to try to collect a reasonably small number of readings which are representative of the large population and in this way we are *sampling the universe of possible readings*. A sample should never be selected but collected randomly. A selected sample is not a representative of the universe and will exhibit higher error of sampling. Errors of sampling are of large consideration in cases where only one or more samples are available as a basis for inductive inference and the total population from which the samples are drawn, is not observable. Obviously we should like that the true mean and the true standard deviation will coincide with the mean and the standard deviation of the sample, which are affected by the following factors:

1. **Size of the sample.** Greater the ratio of the sample number to the population, more nearer will be the sample to the true population. This accuracy increases as the ratio of the square root of those numbers, i.e., if the sample number increases from 10 to 100 then the increase in accuracy will be $\{\sqrt{100}/\sqrt{10}\}$.

2. **Variability of the readings.** When the range of data is shorter, then there is a possibility that the mean of the sample will vary less from the true mean, but in case of a larger range of data, there is a greater possibility of variation of the sample mean from the true mean. So sampling should be as much large as possible otherwise it may give a different representation of the true population.

3. **Frequency distribution of means.** Let μ be the true mean of a normally distributed population and σ the standard deviation. Then

t-TABLE

n*	P=0·9 (90%)	0·8 (80%)	0·7 (70%)	0·6 (60%)	0·5 (50%)	0·4 (40%)	0·3 (30%)	0·2 (20%)	0·1 (10%)	0·05 (5%)	0·02 (2%)	0·01 (1%)	0·001 (0·1%)
1	0·158	0·325	0·510	0·727	1·000	1·376	1·963	3·078	6·314	12·706	31·821	63·657	636·619
2	0·142	0·289	0·445	0·617	0·816	1·061	1·386	1·886	2·920	4·303	6·965	9·925	31·598
3	0·137	0·277	0·424	0·584	0·765	0·978	1·250	1·638	2·353	3·182	4·541	5·841	12·941
4	0·134	0·271	0·414	0·569	0·741	0·941	1·190	1·533	2·132	2·776	3·747	4·604	8·610
5	0·132	0·267	0·408	0·559	0·727	0·920	1·156	1·476	2·015	2·571	3 365	4·032	6·859
6	0·131	0·265	0·404	0·553	0·718	0·906	1·134	1·440	1·943	2·447	3·143	3·707	5·959
7	0·130	0·263	0·402	0·549	0·711	0·896	1·119	1·415	1·895	2·365	2·998	3·499	5·405
8	0·130	0·262	0·399	0·546	0·706	0·889	1·108	1·397	1·360	2·306	2·896	3·355	5·041
9	0·129	0·261	0·398	0·543	0·703	0·883	1·100	1·383	1·833	2·262	2·821	3·250	4·781
10	0·129	0·260	0·397	0·542	0·700	0·879	1·093	1·372	1·812	2·228	2·764	3·169	4·587
11	0·129	0·260	0·396	0·540	0·697	0·876	1·088	1·363	1·796	2·201	2·718	3·106	4·437
12	0·128	0·259	0·395	0·539	0·695	0·873	1·083	1·356	1·782	2·179	2·681	3·055	4·318
13	0·128	0·259	0·394	0·538	0·694	0·870	1·079	1·350	1·771	2·160	2·650	3·012	4·221
14	0·128	0·258	0·393	0·537	0·692	0·868	1·076	1·345	1·761	2·145	2·624	2·977	4·140
15	0·128	0·258	0·393	0·536	0·691	0·866	1·074	1·341	1·753	2·131	2·602	2·947	4·073
16	0·128	0·258	0·392	0·535	0·690	0·865	1·071	1·337	1·746	2·120	2·583	2·921	4·015
17	0·128	0·257	0·392	0·534	0·689	0·863	1·069	1·333	1·740	2·110	2.567	2·898	3·965
18	0·127	0·257	0·392	0·534	0·688	0·862	1·067	1·330	1·734	2·101	2·552	2·878	3·922
19	0·127	0·257	0·391	0·533	0·688	0·861	1·066	1·328	1·729	2·093	2·539	2·861	3·883
20	0·127	0·257	0·391	0·533	0·687	0·860	1·064	1·325	1·725	2·086	2·528	2·845	3·850
21	0·127	0·257	0·391	0·532	0·686	0·859	1·063	1·323	1·721	2·080	2·518	2·831	3·819
22	0·127	0·256	0·390	0·532	0·686	0·858	1·061	1·321	1·717	2·074	2·508	2·819	3·792
23	0·127	0·256	0·390	0·532	0·685	0·858	1·060	1·319	1·714	2·069	2·500	2·807	3·767
24	0·127	0·256	0·390	0·531	0·685	0·857	1·059	1·318	1·711	2·064	2·492	2·797	3·745
25	0·127	0·256	0·390	0·531	0·684	0·856	1·058	1·316	1·708	2·060	2·485	2·787	3·725
26	0·127	0·256	0·390	0·531	0·684	0·856	1·058	1·315	1·706	2·056	2·479	2·779	3·707
27	0·127	0·256	0·389	0·531	0·684	0·855	1·057	1·314	1·703	2·052	2·473	2·771	3·690
28	0·127	0·256	0·389	0·530	0·683	0·855	1·056	1·313	1·701	2·048	2·467	2·763	3·674
29	0·127	0·256	0·389	0·530	0·683	0·854	1·055	1·311	1·699	2·045	2·462	2·756	3·659
30	0·127	0·256	0·389	0·530	0·683	0·854	1·055	1·310	1·697	2·042	2·457	2·750	3·646
∞	0·12566	0·25335	0·38532	0·52440	0·67449	0·84162	1·03643	1·28155	1·64485	1·95996	2·32634	2·57582	3·291

m_1, m_2, etc., the means of the n number of samples, will be normally distributed about μ with a standard deviation $(\sigma/\sqrt{n})$. This is also true for the population not normally distributed but the variations from the normality are not much.

4. **Standard error of mean (standard error).** This frequency distribution of means is of great importance for practical purpose as it indicates that how much the sample mean is varying from the true mean. In other words, it gives an estimate of the error of sampling for which this standard deviation of the true mean is also named as standard error of mean or simply as standard error (e). It should be carefully noted that 'standard deviation' represents commonly the standard deviation of the individual values, and 'standard error' stands for the 'standard deviation of the mean'.

So standard error $=(\sigma/\sqrt{n})$.

Standard error of the sum of two means

$$e_{(m_1+m_2)}=\sqrt{(e_{m_1})^2+(e_{m_2})^2}$$

Standard error of the difference of two means

$$e_{(m_1-m_2)}=\sqrt{(e_{m_1})^2+(e_{m_2})^2}$$

Standard error of product of the two means

$$e_{(m_1m_2)}=\sqrt{(m_1)^2(e_{m_2})^2+(m_2)^2(e_{m_1})^2}$$

Standard error of quotient of the two means

$$e_{(m_1/m_2)}=\sqrt{\frac{(e_{m_1})^2}{(m_2)^2}+\frac{(e_{m_2})^2}{(m_1)^2}}$$

When the true standard error (e) is known then the range, within which the probability (P) of the true mean, may be given by the expression $\bar{x}\pm ke$, where $\bar{x}$ is the mean, k is a suitable constant.

When, P=66 out of 100, then k is approximately equal to 1,
P=95 out of 100, then k is approximately equal to 2,
P=99 out of 100, then k is approximately equal to 3.

The stated probability can also be applied to the range $\bar{x}\pm te_{\bar{x}}$, where factor t has a value which is largest with the smallest sample size and approaches k as the sample size increases, and $e_{\bar{x}}$ is the standard error of the mean.

Now if the arbitrary range be known then probability can be calculated out and in the reverse way, if the probability is fixed (known), the range can be worked out, with the help of the t-table (*vide* page 10.11).

5. **Degrees of freedom.** Degrees of freedom is the number of values of a sample which are freely variable without affecting the mean. For example, say, 3, 5 and 7 are forming a set of sample. So the mean $(\bar{x})$ is equal to 5 and total number of values (n) is 3. Now any two of them can vary freely (i.e., the value may alter) and then the original mean $(\bar{x})$ 5 can be restored by necessary alteration of the third value. So normally the degrees of freedom (n^*) is equal to $(n-1)$ but in more complicated situation n^* may have a lesser value. In the t-table, the t value is given at different probability against the degrees of freedom (n^*) for each set of sample.

6. **Fiducial limit.** With the t value we can estimate the accuracy of the mean of a set of sample. If we choose a value of probability of

occurrence of the true mean to lie in a range on either side of the sample mean and then we can find out the corresponding t values with the degree of freedom. Now we can have the limits with this t value, sample mean and the standard error with the following formula $\bar{x}\pm te_{\bar{x}}$; and this is named as *fiducial* limits of the mean. For example, suppose we want to find out the limits for 95% probability of occurring of the true mean; it can also be stated in the other way that the limit for 5% chance of non-occurrence of the true mean. So here the P=0·05 or (5/100). $n^*=9$ and with reference to the t-table, the t=2·262.

No. of observations	Blood pressure at systole of the heart, mm of Hg.	Deviations $(x-\bar{x})$	Squared deviations $(x-\bar{x})^2$	Working
1	123	0·1	0·01	No. of observations =10 Mean ($\bar{x}$) = 122·9 S.D. = $\sqrt{\frac{\Sigma(x-\bar{x})^2}{(n-1)}}$ $=\pm 4{\cdot}7246$ mm of Hg Standard error = $\frac{\sigma}{\sqrt{n}}$ $=(\pm 4{\cdot}7246/\sqrt{10})$ $=\pm 1{\cdot}49$ mm of Hg
2	130	7·1	50·41	
3	116	−6·9	47·61	
4	118	−4·9	24·01	
5	126	3·1	09·61	
6	124	1·1	1·21	
7	120	−2·9	8·41	
8	128	5·1	26·01	
9	126	3·1	9·61	
10	118	−4·9	24·01	
Total	1229	0	200·90	
Mean	122·9			

The 95% fiducial limits of the mean for the above example are therefore

$$\bar{x}\pm te_{\bar{x}}=122{\cdot}9\pm(2{\cdot}262\times 1{\cdot}49)$$

$$=122{\cdot}9\pm 3{\cdot}46$$

$$=119{\cdot}44 \text{ and } 126{\cdot}36.$$

In other words,

$$\bar{x}-te_{\bar{x}}=119{\cdot}44,$$

$$\bar{x}+te_{\bar{x}}=126{\cdot}36.$$

Fiducial limit of the mean is 119·44 to 126·36.

7. **Probability of a difference.** Just in the reverse way by putting a value of the stated range in the previously calculated fiducial limit's equation of the mean, we can find out the t value. Now with the help of the

t-table we can have the frequency probability that the value is expected to lie. For example, by putting one person's blood pressure (116 mm of Hg) in the above calculated fiducial limits, we may have the *t* value,

$\bar{x}-116=te_{\bar{x}}$, by putting the values of $\bar{x}$, $te_{\bar{x}}$ we can have

$$122{\cdot}9-116=1{\cdot}49\,t$$

$$\therefore\ t=(6{\cdot}9)/(1{\cdot}49)=4{\cdot}63.$$

Now consulting the *t*-table with the *t* value (4·63) and the 9 degrees of freedom, it can be found out that $0{\cdot}01>P>0{\cdot}001$. So by observing the P value in this case, it can be stated that the difference between the expected value and our result (116 mm of Hg) is much for happening due to chance. In other words, there might have some influencing factor which actually carrying this difference. So in such cases we have to revise the expected value or look for some modification of this experimental method adopted for that work.

8 **Significance.** The difference between the experimental results and the control result, may be stated as 'significant' or 'insignificant' by finding the P value. The term 'significance' is used in the sense that it is unlikely to occur due to chance and in the reverse way 'insignificant' means that it is just due to a chance. This significance test is of great importance for the judgement of the scientific investigations. It is customary to regard unlikely to be by chance when $P\leqslant0{\cdot}05$ (i.e., one trial in twenty) as being significant and $P\leqslant0{\cdot}01$ (i.e., one trial in hundred) as highly significant.

9. **Significance test of a difference between two groups of quantitative measurements.** When the same parameter (e.g., blood pressure) is studied in two conditions (e.g., normal and drug treated), then this *t* test can answer whether the change due to the drug treatment is significant or not. We can determine the probability from the *t* value obtained from the following formula:

$$t=\frac{\text{difference between means}}{\text{standard error of difference between means}}$$

i.e., $t=\pm\dfrac{(\bar{a}-\bar{b})}{e_{(a-\bar{b})}}$ when first mean $=a$; second mean $=\bar{b}$.

Standard error of the difference of two means $=\pm e_{(a-\bar{b})}$

or $t=\pm\dfrac{(\bar{a}-\bar{b})}{\sqrt{(e_{\bar{b}})^2+(e_{\bar{b}})^2}}$ as $e_{(a-\bar{b})}=\sqrt{(e_{\bar{b}})^2+(e_{\bar{b}})^2}$

Now let us have an example, where blood pressure is recorded of ten normal persons and ten drug treated persons. The records are stated in the following table from which the significance test is done:

No. of expt.	Blood pressure in mm of Hg		$(a-\bar{a})^2$	$(b-\bar{b})^2$	Working
	Normal (a) (control)	Treated (b) (experimental)			$t=\dfrac{(\bar{a}-\bar{b})}{\sqrt{(e_{\bar{a}})^2+(e_{\bar{b}})}}$
1	225	210	4	81	$=\dfrac{8}{\sqrt{7{\cdot}896+7{\cdot}617}}$ $=\dfrac{8}{\sqrt{15{\cdot}513}}$ $=\pm 2{\cdot}397$ Degrees of freedom $=(n_a-1)+(n_b-1)$ $=(10-1)+(10-1)$ $=20-2$ $=18$
2	220	215	49	16	
3	230	205	9	196	
4	220	220	49	1	
5	240	230	169	121	
6	215	225	144	36	
7	235	210	64	81	
8	220	230	49	121	
9	240	225	169	36	
10	225	220	4	1	
Total	2270	2190	710	690	
Mean	$a=227{\cdot}0$	$\bar{b}=219{\cdot}0$	$e_{\bar{a}}=\pm 2{\cdot}81$	$e_{\bar{b}}=\pm 2{\cdot}76$	
	$(\bar{a}-\bar{b})=8$		$(e_{\bar{a}})^2=7{\cdot}896$	$(e_{\bar{b}})^2=7{\cdot}617$	

Consulting the t-table we find that there is a probability $0{\cdot}05>P>0{\cdot}02$. That means the drug has produced a significant change in blood pressure. If in some other case it appears that the change is not significant but still there is some change then the technique of the experiment may be altered to have a more significant result.

VI. χ^2 test

In many occasions we want to compare two or more groups to find out the efficiency or suitability of the methods, or drugs or something else, i.e., qualitative statistics in which we are not dealing with finer variations but with the quality (attribute) concerned. In such cases the number of individuals in each class is very high but the number of classes is limited or small.

For a single observed value:

$$\chi^2=\frac{(\text{observed frequency}-\text{expected frequency})^2}{\text{expected frequency}}$$

and for a number of values

$$\chi^2=S\left\{\frac{(\text{observed frequency}-\text{expected frequency})^2}{\text{expected frequency}}\right\}$$

Suppose we want to compare the different doses of a vaccine for a particular disease. Three groups are taken, 1st without vaccine, 2nd with low dose vaccine and 3rd with high dose vaccine. Now the cases

Fisher's χ^2-table

n^*	P=0·99	0·98	0·95	0·90	0·80	0·70	0·50	0·30	0·20	0·10	0·05	0·02	0·01
1	0·000157	0·000628	0·00393	0·0158	0·0642	0·148	0·455	1·074	1·642	2·706	3·841	5·412	6·635
2	0·0201	0·0404	0·103	0·211	0·446	0·713	1·386	2·408	3·219	4·605	5·991	7·824	9·210
3	0·115	0·185	0·352	0·584	1·005	1·424	2·366	3·665	4·642	6·251	7·815	9·837	11·345
4	0·297	0·429	0·711	1·064	1·649	2·195	3·357	4·878	5·989	7·779	9·488	11·668	13·277
5	0·554	0·752	1·145	1·610	2·343	3·000	4·351	6·064	7·289	9·236	11·070	13·388	15·086
6	0·872	1·134	1·635	2·204	3·070	3·828	5·348	7·231	8·558	10·645	12·592	15·033	16·812
7	1·239	1·564	2·167	2·833	3·822	4·671	6·346	8·383	9·803	12·017	14·067	16·622	18·475
8	1·646	2·032	2·733	3·490	4·594	5·527	7·344	9·524	11·030	13·362	15·507	18·168	20·090
9	2·088	2·532	3·325	4·168	5·380	6·393	8·343	10·656	12·242	14·684	16·919	19·679	21·666
10	2·558	3·059	3·940	4·865	6·179	7·267	9·342	11·781	13·442	15·987	18·307	21·161	23·209
11	3·053	3·609	4·575	5·578	6·989	8·148	10·341	12·899	14·631	17·275	19·675	22·618	24·725
12	3·571	4·178	5·226	6·304	7·807	9·034	11·340	14·011	15·812	18·549	21·026	24·054	26·217
13	4·107	4·765	5·892	7·042	8·634	9·926	12·340	15·119	16·985	19·812	22·362	25·472	27·688
14	4·660	5·368	6·571	7·790	9·467	10·821	13·339	16·222	18·151	21·064	23·685	26·873	29·141
15	5·229	5·985	7·261	8·547	10·307	11·721	14·339	17·322	19·311	22·307	24·996	28·259	30·578
16	5·812	6·614	7·962	9·312	11·152	12·624	15·338	18·418	20·465	23·542	26·296	29·633	32·000
17	6·408	7·255	8·672	10·085	12·002	13·531	16·338	19·511	21·615	24·769	27·587	30·995	33·409
18	7·015	7·906	9·390	10·865	12·857	14·440	17·338	20·601	22·760	25·989	28·869	32·346	34·805
19	7·633	8·567	10·117	11·651	13·716	15·352	18·338	21·689	23·900	27·204	30·144	33·687	36·191
20	8·260	9·237	10·851	12·443	14·578	16·266	19·337	22·775	25·038	28·412	31·410	35·020	37·566
21	8·897	9·915	11·591	13·240	15·445	17·182	20·337	23·858	26·171	29·615	32·671	36·343	38·932
22	9·542	10·600	12·338	14·041	16·314	18·101	21·337	24·939	27·301	30·813	33·924	37·659	40·289
23	10·196	11·293	13·091	14·848	17·187	19·021	22·337	26·018	28·429	32·007	35·172	38·968	41·638
24	10·856	11·992	13·848	15·659	18·062	19·943	23·337	27·096	29·553	33·196	36·415	40·270	42·980
25	11·524	12·697	14·611	16·473	18·940	20·867	24·337	28·172	30·675	34·382	37·652	41·566	44·314
26	12·198	13·409	15·379	17·292	19·820	21·792	25·336	29·246	31·795	35·563	38·885	42·856	45·642
27	12·879	14·125	16·151	18·114	20·703	22·719	26·336	30·319	32·912	36·741	40·113	44·140	46·963
28	13·565	14·847	16·928	18·939	21·588	23·647	27·336	31·391	34·027	37·916	41·337	45·419	48·278
29	14·256	15·574	17·708	19·768	22·475	24·577	28·336	32·461	35·139	39·087	42·557	46·693	49·588
30	14·953	16·306	18·493	20·599	23·364	25·508	29·336	33·530	36·250	40·256	43·773	47·962	50·892

of occurrence and non-occurrence of the disease are represented in the following table:

	Without vaccine (Placebo)	With low dose vaccine	With high dose vaccine	Total
Attacked with the disease	78	62	60	200
Not attacked	60	68	72	200
Total	138	130	132	400
Expected attacks 50%	69	65	66	
(obs.—exp.)^2	81	9	36	
(obs.—exp.)^2/(exp.)	1·17	0·14	0·55	1·86

It is assumed that if the vaccine has no effect then the 50% of each group should be the expected case number as it is seen in the table that out of total 400, 200 persons are attacked with the disease.

For the above table the χ^2 values for single observation are 1·17, 0·14, 0·55. When these three numbers of χ^2 values are summated together, then

$$\chi^2 = 1{\cdot}86, \text{ i.e., } (1{\cdot}17 + 0{\cdot}14 + 0{\cdot}55 = \chi^2 \text{ for a number of values}).$$

We now have to determine n^*, the degrees of freedom in the sample. It is seen that two of the observed values can alter or vary freely, so that by necessary alteration of the 3rd one, the total 200 may be maintained. So the degree of freedom (n^*) in this case is $(3-1) = 2$. Now by using the χ^2-table the P value may be obtained which in this case is lying in between $0{\cdot}5 > P > 0{\cdot}3$ which means that the above observed results can occur 3 to 5 times in 10 trials. This indicates that the vaccine has a definite or positive protective action though not significant statistically in the above observations which may be statistically significant when is trialed with larger number of observations.

VII. Regression

When we measure two attributes of each of the individuals in a group, e.g., height and weight of a group of human beings, or height of blood pressure and different doses of epinephrine in rats. Now there is possibility of having some correlation between these two attributes. Before considering about the correlation, first of all we have to plot the results graphically by putting controlled variable (epinephrine) as the abscissa and the other as the ordinate. This controlled variable is also known as the independent variable in contrast to the dependent variable. Now if there is any relation between these two variable observations, there will be more or less some regularity in the arrangement of the points; but if there is no such relation between them, the points will be scattered. Now there will have a probability of fitting a straight or curved line (at least approximately) to the points representing somewhat related observations.

Mathematical expression of the best fitting lines. Say, the independent variable be x and y be the dependent variable and when these values are plotted graphically, then the shape of the fitting line will depend on nature of the equation.

Thus the equation $y=a+bx$, when a and b are the intercept on the y axis and slope of the line respectively and x and y are variables. This equation will always represent a straight line. When the value of b is positive the line will slope downwards and to the left and will lie in the upper right and lower left quadrants. If the units for x and y are put to equal lengths then the slope of the line will be given by $b=\tan\theta$, where θ is the angle enclosed by the line and the abscissa on the side of the positive value of x.

That will be the best fitting line for which the sum of deviations of the observed value y from the predicted value Y is zero and the sum of the square of the deviations is minimum (similar to the best or ideal value). It can be shown that the line passing through the mean values of x and y will have the coefficient b of the following value:

$$b=\frac{S\{(x-\bar{x})\ (y-\bar{y})\}}{S(x-\bar{x})^2}.$$

The value of a can be obtained by putting this value of b and values of x and y at a particular point whose co-ordinates are $\bar{x}$ and $\bar{y}$ (since it is known that the line must pass through the mean values of x and y).

The best fitting line so obtained for the observed value of x and y is known as *regression line* relating x and y and the coefficient b is known as the *regression coefficient*. The equation for the best fitting line may also be named as the regression equation.

The calculations may be shortened by using method as that used for shortening the calculation of standard deviation by putting, i.e.,

$$S(x-\bar{x})^2=S(x^2)-\frac{S^2(x)}{n} \text{ and } S(x-\bar{x})\ (y-\bar{y})=S(xy)-\frac{S(x).S(y)}{n}.$$

In some cases we may have a priori reason to know that the best fitting line (regression line) passes through the origin; in other words, the equation of the line is such that when $x=0$ then y is also equal to zero. In such cases the method of determination of the regression equation is still easier as $b=\frac{S(xy)}{S(x^2)}$ and the equation is $y=bx$.

The other formula for regression coefficient is

$$b=r\frac{\text{standard deviation of } y}{\text{standard deviation of } x}$$

when r is the correlation coefficient.

The regression coefficient may be calculated with less complication following the table:

No. of obs.	Epinephrine (μgm) x	% Change (rise) in blood pressure y	Deviations		Product of deviations	Squared deviation	Working
			$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})\times(y-\bar{y})$	$(x-\bar{x})^2$	
1	5	20	−22·5	−74	1665·0	506·25	$b=\frac{S(x-\bar{x})(y-\bar{y})}{S(x-\bar{x})^2}$ $=\frac{7160\cdot0}{2064\cdot5}$ $=3\cdot47$
2	10	35	−17·5	−59	1032·5	306·25	
3	15	45	−12·5	−47	587·5	157·25	
4	20	65	−7·5	−29	217·5	56·25	
5	25	82	−2·5	−12	30·0	6·25	
6	30	102	+2·5	+8	20·0	6·25	
7	35	120	+7·5	+26	195·0	56·25	
8	40	142	+12·5	+48	600·0	157·25	
9	45	159	+17·5	+63	1102·5	306·25	
10	50	170	+22·5	+76	1710·0	506·25	
Total 10	275 $\bar{x}=27\cdot5$	940 $\bar{y}=94\cdot0$	0·0	0·0	7160·0 $S(x-\bar{x})\times(y-\bar{y})=716\cdot0$	2064·50 $S(x-\bar{x})^2=2064\cdot5$	

Now putting the values of b, $\bar{x}$ and $\bar{y}$ (as the line must pass through $\bar{x}$ and $\bar{y}$) in the regression equation $y=a+bx$

$$94\cdot0=a+3\cdot47\times27\cdot5=a+95\cdot42$$

$$\therefore\ a=94\cdot0-95\cdot42=-1\cdot42.$$

So the desired equation for the best fitting line is $y=-1\cdot42+3\cdot47x$ (FIG. 10.7).

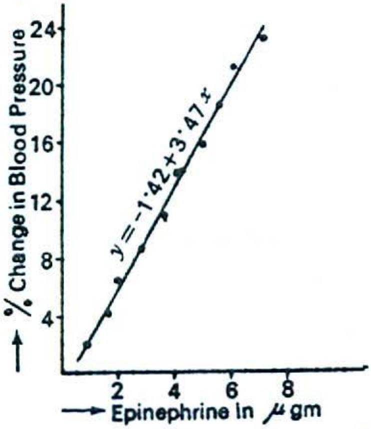


FIG. 10.7. Graphical representation of the regression equation (best fitting line) with the help of the previous table.

Goodness of fitting of a regression line. The goodness of fitting of a regression line can be tested by calculating the squares of the deviations of the *observed values* (x) from the *expected values* (X) and then dividing by the expected values (X), i.e., $\frac{x-\mathrm{X}}{\mathrm{X}}$. Thus values are obtained and are distributed as χ^2. The probability of their occurrence by chance can be obtained by summating them and seeking the value so obtained in a table of χ^2, with, in the case of a straight line, two less degrees of freedom than there are points on the line.

Variance and standard deviation from regression. When the deviations of x are measured from the expected values of X, which is calculated from the corresponding values of y and the regression equation but not from their own mean. Then the sum of squared deviations of x from X divided by the degree of freedom (n^*) will give the variance of x from regression. The number of degree of freedom is determined with the usual method.

Variance of x from regression

$$=S^2_{x\cdot y}=\frac{S(x-\mathrm{X})^2}{(n-2)}.$$

Similarly standard deviation from regression

$$=S_{x\cdot y}=\sqrt{\frac{S(x-\mathrm{X})^2}{(n-2)}}.$$

Standard error of the mean of one dependent variable say, e.g., $\bar{x}$ can be obtained with the following formula $e_{\bar{x}}=\frac{S_{x\cdot y}}{\sqrt{n}}$; and then the significance test of the regression line may be performed.

VIII. Correlation

In the situations where one variable is independent, i.e., in our control and the other is dependent on the former, then we have to consider the relationship with the regression equation. But there are also situations where both the variables are beyond our control andone is dependent or independent on the other. Now to consider whether they have any real or only apparent relationship, we have to go through the term correlation (a modification of regression).

The correlation coefficient (r) is that coefficient which relates x and y when these are measured in standard measure and it also measures the association or correlation between the variables, e.g., children who are above average body weight to be more resistant to a certain disease.

In such cases we can assess the relationship by calculating the r with the following formula:

$$r=\frac{S(x-\bar{x})(y-\bar{y})}{\sqrt{S(x-\bar{x})^2 .\ S(y-\bar{y})^2}}.$$

The alternate formula for correlation coefficient is

$$r=b\,\frac{\text{standard deviation of } x}{\text{standard deviation of } y}$$

where b is the regression coefficient.

The value of correlation coefficient be treated as similar to probability. When $r=1\cdot0$ there is a rigid (definite) connection between the two variables and when $r=0$ there is absolutely no connection or relation.

IX. Analysis of variance

The significance of test—t-test is convenient for quantitative measurements when the data are divisible into two groups. But if the data are of more extensive or of subsidiary divisions then the method known as **analysis of variance** is more convenient, where the t-test is not directly applicable. For example, suppose we want to study the effect of three different ergogenic aids on human subjects of both sexes five male and five female in each group. So with the control or dummy (Placebo) groups, there are four (three experimental and one control) male and four female human subject groups, from which we can calculate the number of mean values. First, the grand mean of whole set of measurements which may be named as $\bar{x}$, secondly, the two mean values of male and female subjects as $\bar{x}_M$ and $\bar{x}_F$, and thirdly, mean values of the four groups comprising of both male and female subjects—$\bar{x}_A$, $\bar{x}_B$, $\bar{x}_C$ and $\bar{x}_D$. The last set of mean values can be obtained from the four male and four female groups separately as $\bar{x}_{AM}$, $\bar{x}_{BM}$, $\bar{x}_{CM}$, $\bar{x}_{DM}$ and $\bar{x}_{AF}$, $\bar{x}_{BF}$, $\bar{x}_{CF}$, $\bar{x}_{DF}$ respectively.

Now if we put the working hypothesis as that the subjects are devoid of any effect due to the ergogenic aid or sex. Then all the means calculated are the estimates of true mean μ of the population from where the sample is collected, with variance $(\sigma/\sqrt{n})$, when $\sigma=$true standard deviation of the population and $n=$number of subjects in a particular estimate of the mean. So all the calculated variances are the estimates of the true variance and the difference in these two will be the effect of random sampling. The ratio of two estimates of variance of a normally distributed population is denoted as the statistic F or e^{2z} (here e is the base of natural logarithms and not the standard error). Similarly the distribution is known as the statistic z, which is half the difference between two natural logarithm of two estimates of the variance. The values of e^{2z} and z for common values of n (size of the samples) and the corresponding value of P (probability can be obtained from the table of e^{2z}. Similarly the probability of ratio of the variances of grand mean $(\bar{x})$ and this group of males can be determined from the table of e^{2z}. The value of P will indicate whether the results are drawn from a single normally distributed population

Example: *If* m =number of groups, n =number of individuals in each group, $\bar{x}$ =grand mean, $\bar{x}_L$ =general term for group mean, $\bar{x}_S$ =general term for sex mean and x_1, x_2, x_3, etc., i.e., x are the individual values. Then the following table will show the method of calculation of variance for the above mentioned set of data:

Item numbers	Items (included in the calculation of variance)	Degree of freedom	Sum of the squared deviations	Variance $\left(\frac{S(d)^2}{(n-1)}\right)$
1	Between groups	$m-1$ $4-1=3$	$S(\bar{x}-\bar{x}_L)^2$ (between grand mean and treatment means)	$\frac{S(\bar{x}-\bar{x}_L)^2}{(m-1)}$
2	Between sexes	$2-1=1$	$S(\bar{x}-\bar{x}_S)^2$ (between grand mean and sex means)	$\frac{S(\bar{x}-\bar{x}_S)^2}{1}$
3	As a whole	$mn-1$ $4\times10-1$ $=39$	$S(x-\bar{x})^2$ (between grand mean and individual values)	$\frac{S(x-\bar{x})^2}{mn-1}$
4	Within groups	$m(n-1)$ $4(10-1)$ $=36$	$S(\bar{x}_L-x)^2$ (between treatment means and individual values)	$\frac{S(\bar{x}_L-x)^2}{m(n-1)}$
5	Residual (within groups)	$(mn-m-1)$ $40-4-1$ $=35$	$S(x-\bar{x})^2-\{S(\bar{x}-\bar{x}_L)^2+S(\bar{x}-\bar{x}_S)^2\}$ $=S(x-\bar{x}_L)^2-S(\bar{x}-\bar{x}_S)^2$ (by difference, item 3. —1. —2.	$\frac{S(x-\bar{x}_L)^2-S(\bar{x}-\bar{x}_S)^2}{mn-m-1}$

X. Experimental design

Under the previous headings we have considered how to analyse or interpret the experimental results when they have been obtained. But this analysis or interpretation of the results or observations depends mostly on the method or way by which they are obtained. Because it is quite possible that the observations, obtained out of a long laborious experiment, may not have any useful information. So it is preferred that the question of design of experiment and analysis of observations should be considered at the same time. The analysis of observations has already been considered in the previous pages, here we have to consider the experimental design only. During consideration of the experimental design the following factors are to be kept in mind:

1. **Range of reliable induction.** The design of the experiment depends upon the type of problem we want to study and the type of inference we wish to make. To have a more general inference the sample (observation) should be taken from a hypothetical infinite population and the method of sampling would have to afford a random sample

and not one biased to suit the convenience of the experiments. Say for example, it is observed that the administration of epinephrine causes rise of blood pressure in cats, dogs and rabbits; from this the inference may be drawn that the epinephrine causes rise of blood pressure in all mammals. This inference may be correct but to justify the inference, the observations should be performed on more number of mammalian group of animals which will make the interence more reliable.

2. **Null hypothesis.** When some inferences are drawn on some common features present in the observed facts—then the inference drawn is named as **hypothesis.** During setting up of the experimental design, one control and another experimental groups always be set up, i.e., one group without the factor X (control) and other group with the factor X (experimental). Now during analysis the statistical tests are always concerned with the difference between the observations (control and experimental) and not with the isolated observation. So the control observation is equally important along with the experimental one. By considering the difference in observations, all other factors except the factor X are eliminated. So the result obtained is expected to be due to the factor X only. There is one hypothesis that two or more sets of measurements are likely to have drawn by chance from the same population. This can well be tested with the statistical methods. So according to this hypothesis the observed differences are likely to be due to errors of sampling and there is not necessarily any confirmable difference between the groups. This hypothesis is named as the **null hypothesis.** This is specially suitable for rigid deductions and practical use.

3. **Segregation of causes of variation.** During consideration of the experimental design we should be justified that the results or differences obtained in the group-containing factor X, is due to the presence of the factor X only and in that case only. We can attribute the function (effect) to X. But it is inevitable that a large number of other factors (known or unknown) will effect the experimental subjects, e.g., age, height, weight, physical condition, etc. And we cannot say that these small or minor differences will not affect the observed results. So we should have an eye that any small variation in between the subjects is to be avoided as far as practicable. The experiment must be so performed that the non-specific factors may affect the result contributing to the residual variance but not to the variance related to the factor X. This may be achieved by random allocation.

4. **Random allecation.** The selection of the experimental animals (or human subjects) should be made at random, i.e., all the weaker or stronger animals should not be placed in the same group but the groups should contain the animals of both weak and strong groups equally. This may be done by numbering the individual animal and the numbered cards were shuffled well. Then by pulling one card at a time at random and putting that numbered animal to one group, the purpose of random allocation may be served.

5. **Reduction of experimental error.** When all the above factors are considered in an experimental design, the significance of the results will depend directly on the magnitude difference produced by the factor X (i.e., difference between the experimental and the control), on the number of observations and inversely on the standard deviation or error.

The first factor, i.e., the difference produced by the factor X is beyond our control yet in medical experiments this may be affected by the alteration of the dosage and duration of the drug. For a more reliable inference, the number of observation should be as high as possible. The experimental material should be made as homogeneous as possible and experimental conditions should be rigidly controlled and statistical analysis should be performed to estimate the cause of variation not attributable to the experimental factors. This may be summarised in the following headings: (*a*) **Replication**—The experiment should be conducted repeatedly in identical condition. This will reduce the experimental error and give precise to the observation. (*b*) **Error control**—Any factor, e.g., age, sex, temperature, etc., which may affect the experimental observation should be considered strictly. The experimental and the control experiment should be performed on the same condition from every aspect. (*c*) **Randomisation**—By selecting the subject, the validity to the data may be increased.

XI. Nomogram

It is the graphical representation of mathematical laws which are expressed analytically by means of equations. The term is sometimes restricted to a special type of chart which is used by bringing the points of these scales into alignments. The theory of nomogram is based mostly on analytical geometry. If the nomogram is once prepared with some labour, complicated problems of that system can be solved with speed and with slight labour, though the solutions are not of high accuracy. This is particularly helpful when many similar numerical problems are to be solved. Say for a simple example, it is well known that the height, weight and body surface of a person have got some definite relationship, so when the nomogram is prepared with that definite relationship, the value of anyone of these three parameters can be solved out from the remaining two (Figs. 10.9 & 10.10).

The nomograms of equations of many variables are prepared by using a sequence of scale alignments or by employing networks of scales. They can be used by a person of ordinary knowledge or experience. These graphic representations are used widely in the problems of engineering, industry, natural and physical sciences.

In the more general forms of nomogram, the alignment chart for the solutions ranged in various ways, and one or more of the scales may be curved. The nature of the resulting scale-curved or straight depends upon the parametric equations. Only three scales need be drawn and they can be used by interpolating a straight line over these scales, i.e., interpolating the straight line drawn through the two known points ($X_1=6$, $Y_1=8$ as Fig. 10.8), corresponding third (unknown) value ($Z_1=10$, Fig. 10.8)—point of intersection—can be obtained. Now in actual practice if these lines are really drawn, a few lines will mar the graph. This can be avoided by stretching a fine thread or by

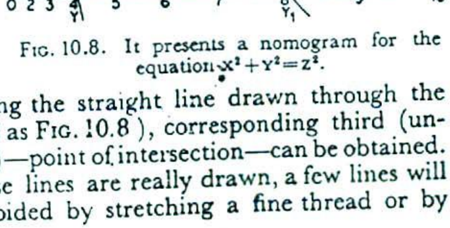


Fig. 10.8. It presents a nomogram for the equation $X^2+Y^2=Z^2$.

placing a straight line of a transparent scale in position across the chart (Fig. 10.8).

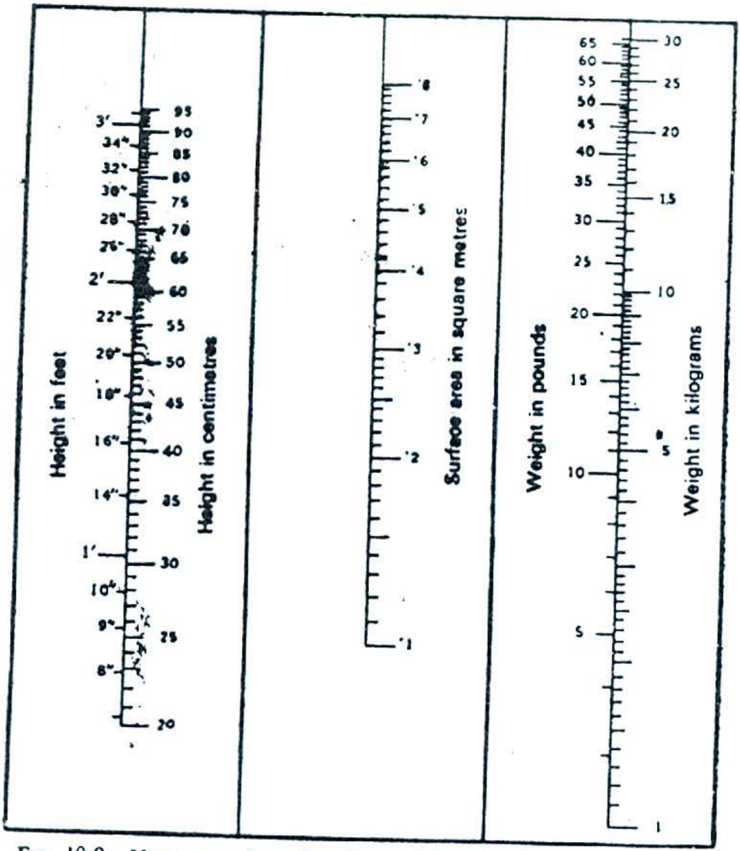


Fig. 10.9. Nomogram for estimating surface area of infants and young children. To determine the surface area of the individual it can be drawn a straight line between the point representing his height on the left-hand vertical scale to the point representing his weight on the right-hand vertical scale. The point at which this line intersects the middle vertical scale represents the individual's surface area in square metres.

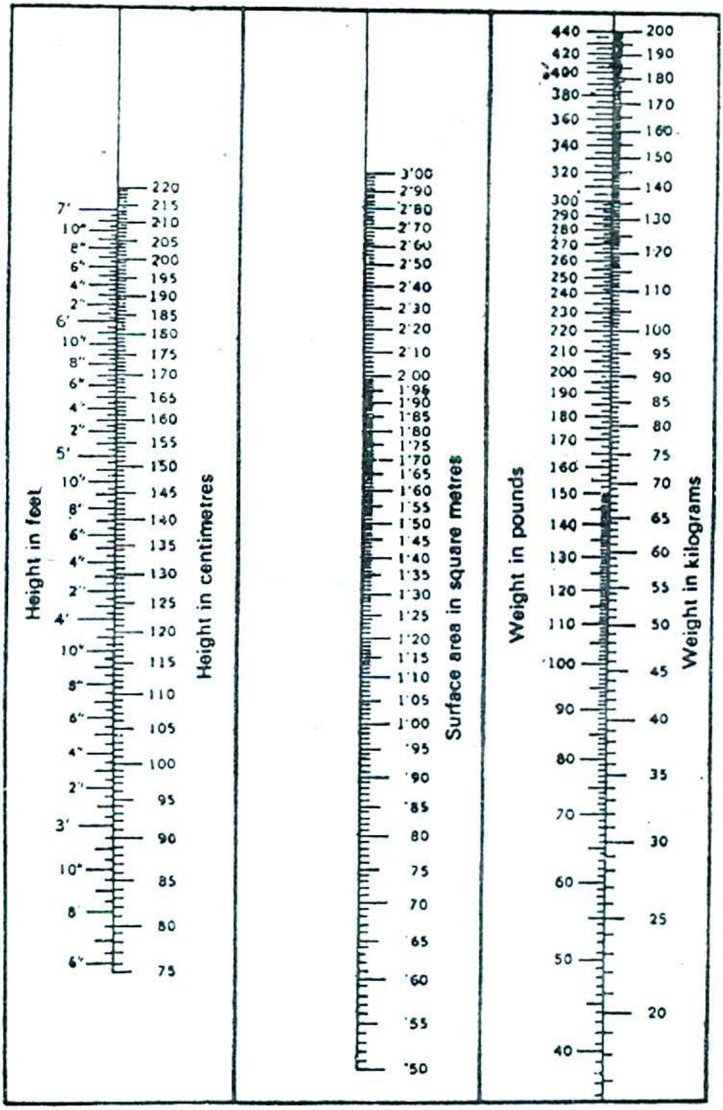


FIG.10.10. Nomogram for estimating surface area of older children and adults. The description is same as in FIG. 10.9.