Introduction

Statistical methods are being increasingly used in new fields of scientific work and practical affairs with the dual objectives of economic gain, and of improvement of the validity of data and its conclusions (Mahalanobi, 1965).

Modern statistical methods, according to Miller and Wichern, 1977) are, concerned with the generation and analysis of numerical information or data. These methods are perhaps best appreciated in the context of scientific research.

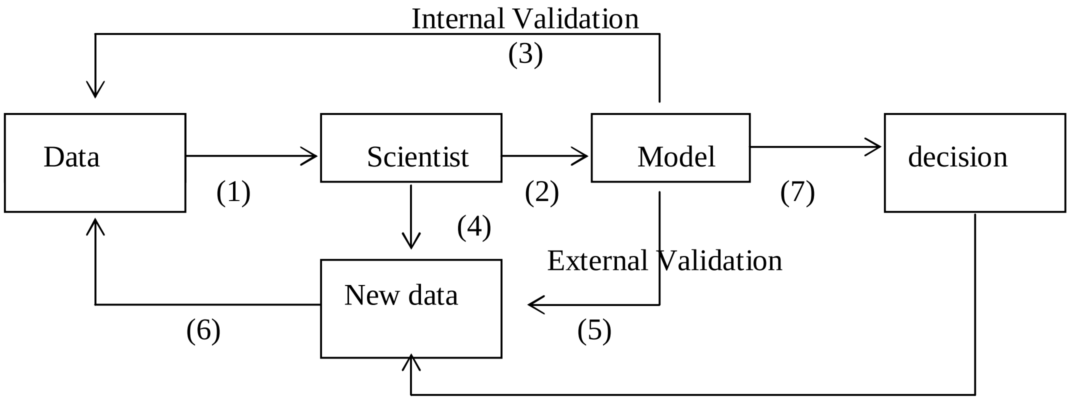
The goal of any research effort is to gain, understanding of observable Phenomena. Given this understanding, a further goal may be to predict or control these and products of these phenomena. For one thing, scientific method places emphasis on gaining knowledge through the process of observation, whatever is said about behaviour is reasoned from systematic observation and is tasted and retested by observation.

However, drawing general conclusions from experimental data (induction) is inherently, an uncertain activity and scientist rarely conclude that a proposition is true; rather they speak in terms of degree to which propositions agree with the observed facts.

Part of the task of designing a study involves deciding before hand which statistical procedures to be used and what criteria will be involved in reasoning from the statistical results back to the population under study. The selection of statistical models depends of course on what kinds of mathematical deductions the research plan calls for. These involves, consideration, among other things, of what population, characteristics are of interest (e.g. averages, dispersions etc), what population comparisons might be made, and what type by measurement scale is involved.

A model may range from a very simple to a very complex set of propositions and the testing of a model.

A paradign of the evolution of scientific knowledge is as follows:



Referring to the figure, we make the following comments:

(1) Data may consist of causal observations, measurements collected by other investigators, or the results of carefully planned experiments or surveys.

(2) After observing the data, the scientist formulates proposition or hypothesis purporting to explain the particular outcomes observed. This set of propositions constitutes the model of the phenomenon, of interest and is the creation of the scientists imagination from the model certain consequences may be logically deduced. These consequences are also part of the model.

(3) The model must yield consequences that are consistent with the already observed facts. The determination that these consequences are indeed consistent with the data in hand may require analysis. We call this step internal model validation.

(4) When satisfied within the internal model validation, the next question to ask is whether the predictions made using the model are consistent with the new data. At this point the investigator must plan and carry out an experiment that will yield new data. These data are then used to evaluate model performance.

(5&6) The comparism of model predictions with new data is called external model validation. If the model appears to be inadequate (does not fit the facts), modifications of the model will be made. These revised model is validated using all data currently available. If the model seems to be making reasonable predictions, it is accepted tentatively as a working hypothesis. The cycle repeats if at any point. The hypothesis is inconsistent with new data.

(7&8) When a model is accepted as a working hypothesis, it may be shared through publication or private communication. The results produced by an acceptable model are ordinarily used by planners, forecastors, policy makers, to make decisions.

It is this feature that makes statistical activity so fundamental to scientific inquiry. Statistics provides guidelines for efficiently generating informative data and for constructing tests of these data against certain propositions. We might classify the study of statistics as the study of variation or differences in data. We can divide the sources of variation into three man categories

(a) Inherent variation

(b) Experimentally induced variation

(c) Variation due to errors or mistakes.

Inherent variation is made and it varies from individuals to individuals. This variation is inherent in the collection of individuals as characteristics such as age, height, size of family, etc.

Experimentally induced variation are divided into control and experimental groups. This variation are due to treatment effect on the subject.

Variation due to errors or mistakes in handling data. This kind of variation can be minimized by careful data processing.

The scope of statistics extends over all disciplines and is essentially an applied science. Its only justification lies in the help it can give in solving a problem. Its aim is to reach a decision, on a probabilistic basis, on available evidence. According to mahalanobis, "if science is based on observation and measurement and if each set of measurements is a statistical sample, then all scientific conclusions must be of the native of uncertain inference with in principle, a known margin of uncertainty.

Statistical Control of Measurement Processes

Measurement has been defined as the assignment of a number to characterize some property of an object.

Eisenhart (1962) asserts measurement is the assignment of numbers to material things to represent the relations existing among them with respect to particular properties. One aspect of statistics concerns how observations are reported in terms of measurement.

Measurement involves the application of some numeric or symbolic scheme used to designate characteristics of a variable. Measurement bridges the gap between what we record as observations of a variable in the real world and what we may define as a variable in a statistical model. Different measurement schemes, include the nominal, a designation of such classes of characteristics; the ordinal, sub-classigications that are rank ordered; the interval, ranking with known intervals but with an arbitrary zero; and the ratio ranking with known intervals but with a true zero.

Accuracy in measurement is like accuracy in language and it depends on the user. Two facets of the measurement problem concern the qualities of validity and reliability.

Validity is the degree to which the researcher measures what they claim to measure. The question of validity is a "question of goodness of fit" between who the researcher has defined as a characteristic of a phenomenom and what he or she is reporting in the language of measurement.

Reliability is the external and internal consistency of measurement. Traditionally, the quality of statistical data refers to its accuracy. A measurement system that produces accurate data in the sense of high precision and low bias is one of high quality.

By accuracy scientists means the degree of the agreement of the measurement to the true value. In effect accuracy refers to closeness of measurements to the truth and precision to closeness of measurements to each other, while Bias refers to the measure of the difference between the true value and the limiting mean or the average of the measurements over all possible repetitions of the process.

Although uncertain events can not be predicted with unerring accuracy, they are often observed to follow certain regular patterns or laws. Probability theory is the result of man's attempt to discover the laws of chance.

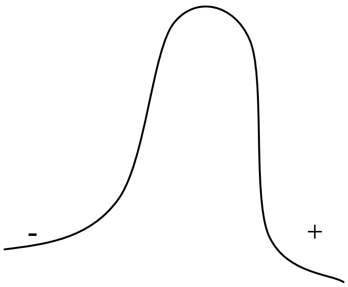
Statistical inferences share with the activities as per the element of uncertainty. Statistical inferences are generalizations on the basis of limited information and hence subject to unpredictable amounts of error.

In fact, a fundamental process in statistics is the attempt to evaluate using probability theory, the precision (or limits on the error) with which an inference can be made in a given set of circumstances. This process naturally leads to the comparison of competing statistical techniques and the development of recommendations for or against using a given technique in a given situation. The basic element in the theory is the probability distribution. The most important and widely used distribution in statistical analysis in the Normal (or Gaussian) distribution. The distribution was developed by Demoivre, Laplace and Gauss separately as a mathematical displacement distribution. The curve represents the distribution of values from the true value (mean). This mathematical concept is used widely today and is referred to as the "law of Errors".

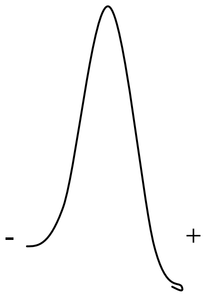
The law of errors tells us that any group of measurements on the same sample, subject, or homogeneous group normally appears clustered about the mean observation. This distribution shows an astonishing degree of regularity in errors of measurement. Many standard unvariate statistical methods are based on the assumption that data are normally distributed.

The normal pdf is symmetric about its mean μ and extends from $-\infty$ to $+\infty$. The frequency of the distribution falls about the mean in a some what predictable bell shaped curve often referred to as a Gaussian distribution. Any normal curve will be bell-shaped and symmetrical round its mean, but its actual form (height and width) will differ according to the value of its mean μ and its standard deviation $\sigma_{\bar{x}}$.

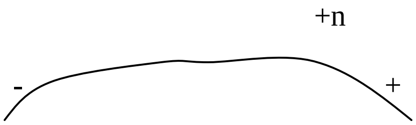
The Normal curve has an important characteristic. No matter what its shape, i.e. broad or narrow, the area beneath the curve is distributed in a particular way. The normal distribution can resemble a tall single peaked narrow curve or a flatish long tailed broad unmodel i.e a simple peak and symmetrical about that peak. The shapes of distributions can take the following form.



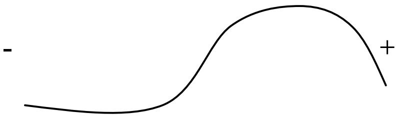
(A) Bell shaped



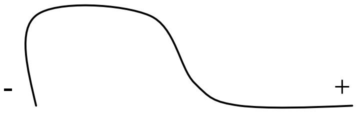
(B) Peaked



(C) Flat



(F) Negative skew



(E) Positive skew

When distributions can be approximated to the normal distribution curve, we have at least, one type of basis for making probability statements about. Parameters and likely sampling error of statistics. Instead of visualizing the normal curve as an area divided into parts we consider it as a large collection of sample means. Clearly, the means of all the sample means is the best estimate was have of the mean of the population from which the samples have been drawn.

Normal distribution curve can then be defined as a particular functional relationship between deviations, about the mean of a distribution and the probability of these different deviations occurring.

The normal curve is a theoretical distribution only. We may never find sample distributions or population distributions that precisely fit the normal curve, but in many cases we do not go very far wrong if we assume that certain population have this kind of distribution.

Assuming that a population has this distribution then implies that a random sample of that population will also have a normal distribution. Additionally, the sampling distribution of a certain statistics can also be assumed to have a normal distribution e.g. the sampling distribution of the mean in a normally distributed population.

In summary, what we have then, in the normal distribution curve is simply a model of a distribution having certain relations between points along the baseline (expressed in σ units) and probabilities of occurrence (areas, in segments of the curve). This point is mentioned because it is sometimes erroneously assumed that a normal distribution is always bell shaped. The key feature of the normal distribution is the defined relation between each σ unit and respective areas under the curve.

The essence of statistical methods is to reason from characteristics of samples (statistics) in order to estimate characteristics of populations (parameters) we are in a position to employ the logic of statistical inference when samples have been randomly drawn from populations. Sample characteristics are incorporated in a sample distribution, whereas a population distribution, incorporates the population characteristics that are to be estimated. A sampling distribution characterizes the likely deviation of a given statistic about a parameter. This sampling distribution have as its standard deviation a standard error of the standard deviation, a concept analogous to the standard error of the mean.

The sampling error is an estimate of how statistics may be expected to dviate from parameters when sampling randomly from a given population.

In practical terms, whereas random sampling will yield sample charactistics that tend toward the population characteristics, we cannot expect sample characteristics to be precisely the same as the population characteristics.

The laws of chance allow sample characteristics to deviate from population characteristics, but our knowledge of these laws allow us to estimate what kinds of deviations to expect. This is the essence of statistical method and the underlying logic of sampling statistics.

The mathematical basis of statistical inference is probability theory.

Probability theory is the result of man's attempt to discover the "laws of chance" statistical inferences are generalizations on the basis of limited information and hence subject to unpredictable amount of error. In fact, a fundamental process in statistics is the attempt to evaluate. Using probability theory, the precision (or limits on the error) with which an inference can be made in a given set of circumstances.

The basic element in probability theory is the probability distribution. Probability distribution is formed when probabilities are associated with numbers. Probability ordinarily is defined in the context of performing an “experiment”.

Experiments frequently yield numerical outcomes. A ‘variable’ that assumed the values of the outcomes of an experiment is called a random variable.

Random variables may be either discrete or continuous. If the set of all possible values of the random variable is either finite or countably infinite, then the random variable is discrete, and its probability structure is described by the probability mass function (p.m.f.) let x be the random variable of interest in an experiment. Suppose x is that a couple has three children. The set of all possible values of the random variable x is finite and the random variable is discrete and its probability function is denoted as P(x).

However, if the set of all possible values of x is an interval, say the height or weight of students, then the random variable is continuous and its probability structure is described by probability density function (or pdf) denoted as f (x) are called probability distributions.

The mean of a probability distribution is a measure of its control tendency or location. We define the mean $(e.g.\mu)$ as

$$\mu = \sum_{i=1}^{n} \times P(x) \quad \text{for } x \text{ discrete}$$

or

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \quad \text{for x continuous.}$$

We may also express the mean in terms of the expected value or long run average value of the random variable x as

$$\mu = \mathrm{E}(\times) = \sum \mathrm{P}(\times) \qquad \text{for x discrete}$$

or

$$\mu = \mathrm{E}(\times) = \int_{-\infty}^{\infty} \mathrm{P}(\times) d \times \text{for x continuous}$$

Where E denotes the expected value operator.

The spread or dispersion of a probability distribution can be measured by the variance, defined as

$$\sigma^2 = \sum_{i=1}^{n} (\times - \mu)^2 \mathrm{P}(x) \qquad \text{for x discrete}$$

or

$$\sigma^2 = \int_{-\infty}^{\infty} (\times - \mu)^2 f(x) dx \qquad \text{for x continuous}$$

Note that variance can also be expressed entirely in terms of expectation, since

$$\sigma^2 = \mathrm{E}\left[(\times - \mu)^2\right]$$

Finally, the variance is used so extensively that it is convenient to define a variance operator Var (or V) such that

$$\text{Var } (\times) = \text{V}(\times) = \mathrm{E}\left[(\times - \mu)^2\right] = \sigma^2$$

The concepts of expected value (E) and variance (v) are used extensively in statistics. It may be helpful to review several elementary results concerning these operators.

If x is a random variable and C is a constant, then

(i) $\mathrm{E}(C) = C$

(ii) $\mathrm{E}(\times) = \mu$

(iii) $\mathrm{E}(C\times) = C\mathrm{E}(\times) = C\mu$

(iv) $V(C) = 0$

(v) $V(\times) = \sigma^2$

(vi) $V(C\times) = C^2V(\times) = C^2\sigma^2$

If there are two random variables for example $\times_1$ with $\mathrm{E}(\times_1) = \mu_1$and $V(\times_1) = \sigma_1^2$ and $\times_2$ with $\mathrm{E}(\times_2) = \mu_2$ and $V(\times_2) = \sigma_2^2$, then we have

(vii) $\mathrm{E}(\times_1 + \times_2)$ = $\mathrm{E}(\times_1)$ + $\mathrm{E}(\times_2)$ = $\mu_1 + \mu_2$

(viii) $V(\times_1 + \times_2)$ = $V(\times_1)$ + $V(\times_2)$ + 2 cov $(\times_1, \times_2)$

Where

cov$(\times_1 + \times_2)$ = E[$(\times_1 + \mu_1)$ $(\times_2 - \mu_2)$]

is the covariance of the random variable $\times_1$ *and* $\times_2$. The covariance is a measure of independence; that is, if $\times_1$ *and* $\times_2$ are independent, then cov ($\times_1$ *and* $\times_2$)= 0. we may also show that

(ix) $V(\times_1 - \times_2) = V(\times_1) + V(\times_2) - 2\,\mathrm{cov}(\times_1, \times_2)$

If $\times_1$ *and* $\times_2$ are independent, then we have

(x) $V(\times_1 - \times_2) = V(\times_1) + V(\times_2) = \sigma_1^2 + \sigma_2^2$ and

(xi) $\mathrm{E}(\times_1 - \times_2)$ = $\mathrm{E}(\times_1)$ - $\mathrm{E}(\times_2)$ = $\mu_1 + \mu_2$

Note that

(xii) $\mathrm{E}\frac{(\times_1)}{\times_2} \neq \frac{\mathrm{E}(\times_1)}{\mathrm{E}(\times_2)}$

Regardless of whether or not $\times_1$ *and* $\times_2$ are independent.

Sampling and Sampling Distributions

The objective of statistical inference is to draw conclusions about a population using a sample from that population.

That is, if the population contains N elements, and a sample of n of them is to be selected. Then if each of the $\mathrm{N}!\backslash(\mathrm{N}-n)!n!$, possible samples has an equal probability of being chosen, the procedure employed is called random sampling. In practice, it is difficult to obtain random samples, and table of random numbers may be helpful.

Statistical inference makes considerable use of quantities computed from the observations in the sample. Statisticians define a statistic as any function of the observation in a sample that does not contain unknown parameters.

e.g suppose that $\times_1, \times_2, - - - \times_n$ represents a sample, then the sample mean.

$$\overline{\times} = \frac{\sum_{i=1}^{n} \times_i}{n}$$

And the sample variance

$$S^2 = \frac{\sum_{i=1}^{n} \left(\times_i - \overline{\times}\right)^2}{n}$$

are both statistics.

These quantities are measures of central tendency and dispersion of the sample respectively. Sometimes, $S = VS^2$, called the sample standard deviation is used to measure dispersion.

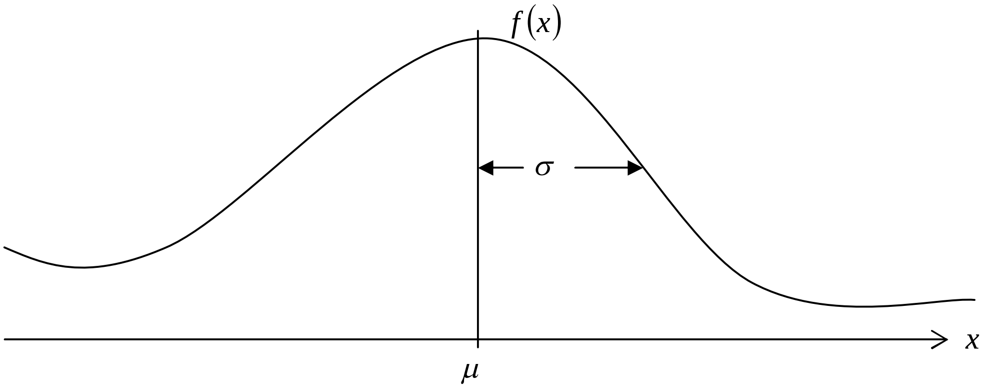
Often use determine the probability distribution of a particular statistic. If the probability distribution of the population from which the sample was drawn is known. The probability distribution of a statistic is called a sampling distribution. We will discuss several useful sampling distributions.

One of the most important sampling distributions to the normal distribution.

Normal Distribution

The normal (or Gaussian) distribution is the most important and widely used distribution in statistical analysis.

The reason is that many every day issues appear to be approximately normally distributed. Another reason for the use of normal distribution is that many commonly used statistics have approximate normal distributions. The normal distribution is symmetric about its mean μ and extends from $-\infty$ to $+\infty$ as shown below



The random variable × has a normal distribution if its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2x}} \ell^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$-\infty < x < \infty$$

Where $-\infty < \mu < \infty$ is the means of the distribution and $\sigma^2 > 0$ is the variance. We always use the notation $\times n\mathrm{N}(\mu, \sigma^2)$ to denote that x is distributed normally with mean μ and variance σ^2. The normal distribution plays a central role in statistical methodology.

An important special case of the normal distribution is the standard normal distribution. This is a normal distribution that has mean $(\mu = 0)$ and variance $\sigma^2 = 1$.

We see that if

$\times \sim \mathrm{N}(\mu, \sigma^2)$ then the random variable

$$\mathrm{Z} = \frac{\times - \mu}{\sigma}$$

Follows the standard normal distribution denoted as $\mathrm{Z} \sim \mathrm{N}(0,1)$.

The pdf of the standard normal distribution Z is given by

$$f(z) = \frac{1}{\sqrt{2\overline{\wedge}}} \ell^{-\frac{z^2}{2}} \qquad -\infty < z < \infty$$

Thus an observation x from a population with mean μ and standard deviation σ, has a z score or z value defined by

$$\mathrm{Z} = \frac{\times - \mu}{\sigma}$$

A z score measures how many standard deviations an observation is to the right or left of the mean. Since σ is never negative, a positive x score measures the number of standard deviations an observation is above the mean, and a negative z score gives the number of standard deviations an observation is below the mean.

Properties of the normal curve

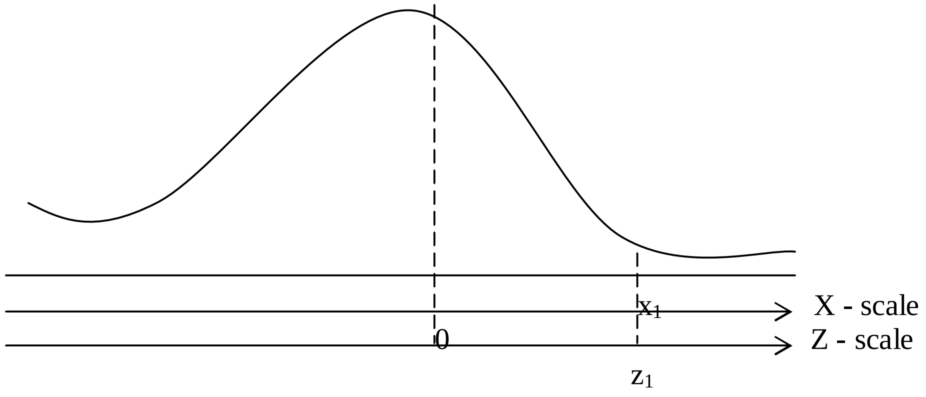
1. It is bell shaped.
2. It is symmetrical i.e Mean – Median – Mode.
3. Area under the normal distribution shows the deviation between the set of observations.
4. the set of observations before and after the mean are equal.
5. when a sample size is large the distribution tends to normal.

Areas Under the Normal Curve

Whenever a problem involving normal distribution is to be tackled, a diagram should be drawn. This makes the situation clearer and it is essential in an exam.

The normal distribution, has to be coverted via the transformation of z score to a probability.

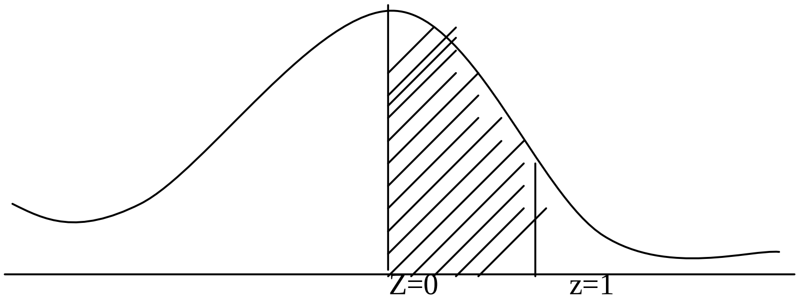
Consider the following diagram under which we are to transform an x- scale (the scale of the original variate) to a z-scale (the scale of the new variate.



Note $Z = \frac{\times - \mu}{\sigma}$

Since the area under the curve add up to 1, the area under the curve on either side of the mean must be equal to ½ .

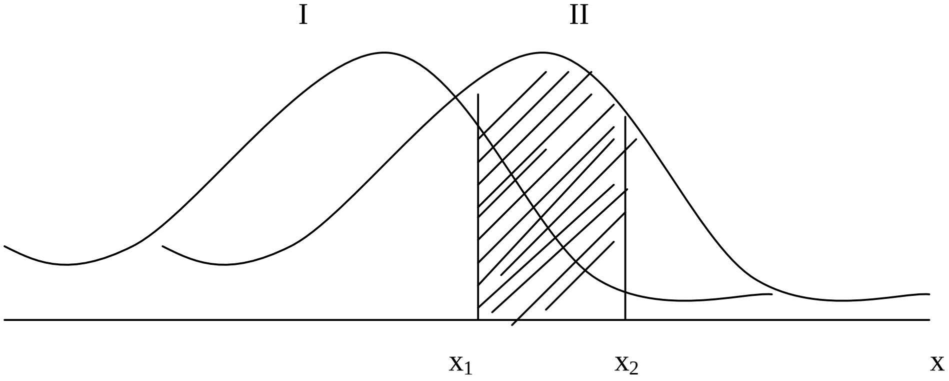
Normal distribution tables have been tabulated from z = 0 to z = i as below:



All probabilities can be calculated from the tabulation if it is remembered that the distribution is symmetrical and that the area to the left of the mean is equal to ½ i.e the probability that z $\leq$ 0 is equal to ½.

When x is between the values $x = x_1$ and $x = x_2$, the random variable z will fall between the corresponding values

$$Z_1 = \frac{x_1 - \mu}{\sigma} \quad \text{and} \quad Z_2 = \frac{x_2 - \mu}{\sigma}$$



For the normal curve, $\mathrm{P}(x_1 < x < x_2)$ is represented by the area of the shaded region.

Hence we have

$$\mathrm{P}(x_1 < x < x_2) = \mathrm{P}(z_1 < x < z_2)$$

Normal Distribution

For many random variables, the probability distribution is a specific bell shaped curve called the normal curve or Gaussian curve. It is the most useful probability distribution in statistics. For instance, errors made in measuring physical and economic phenomena often are distributed normally. The normal distribution is a continuous random variable with parameters $\mathrm{N}(\mu, \sigma^2)$, where $\mu \mathrm{E} \Re, \sigma^2 \rangle 0$.

It has the pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\barwedge}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} x E \Re$$

a) Standard Normal Distribution

A random variable Z is called standard normal if its probability distribution is

$$f(z) = \frac{1}{\sqrt{2\barwedge}} e - {}^{1}\!/_{2}\, z^2$$

The constant $\hat{\theta}\ {}^{1}\!/_{\sqrt{2\barwedge}}$ is a scale factor required to make the total area 1. The symbolic $\barwedge$ and e denotes important mathematical constants, approximately 3.14 and 2.72 respectively.

The standard normal random variable is an important version of a $\mathrm{N}(\mu, \sigma^2)$ random variable. The symbol for the $\mathrm{Z} = (\times - \mu) \backslash \sigma$ and Z has the standard normal pdf with

$\mathrm{N}(O,1)$. This relationship between X and Z allows one to compute probabilities for X.

Therefore, if Z is $\mathrm{N}(O,1))$ and C is any point

$$P\sigma(Z \geq C) = P\sigma(Z > C)$$

Hence if X is called a standard normal random variable, then for any a < b.

$$P\sigma(a < \times \leq b) = \int_a^b e^{-x^2/2} \frac{dx}{\sqrt{2\overline{\wedge}}}.$$

The condition that $P(-\infty < \times < \infty)$ follows from the classical integral.

$$\int_{-\infty}^{\infty} e - x^2/2 \, dx = \sqrt{2\overline{\wedge}}\ .$$

Standard normal $\sigma.v.s$ are a sort of universal random variables in probabilistic modelling. This become clear through the central limit theorem (CLT).

b) Functional Central Limit Theorem

The central limit theorem says that the properly normalized and centred partial sums of an independent identically distributed finite variance sequence. Converge in distribution to a normal distribution. The central limit theorem is actually a mathematical statement about the standardized version of a sampling distribution from Normal Populations.

c) Sampling from Normal Populations

the normal distribution is an approximation to the sampling distribution of means. The mean and variance of the sampling distribution of means give many insights into the behaviour of mean ($\bar{x}$) as an estimate of μ. They show that $\bar{x}$ has the correct aim for estimating μ, and they give some general impressions concerning the amount by which $\bar{x}$ can be expected to vary about the population mean under differing circumstances. The attainment of these goal requires probability statements about $\bar{x}$, and such statements can come only from a reasonably exact specification of the entire sampling distribution.

Many statistical problems involve the estimation of population means, and $\bar{x}$ is a reasonable estimator of the population mean. An important property of sample averages is that their sampling distributions can often be approximated by a normal distribution regardless of the form of the parent population. This result is known as the central limit theorem (CLT).

CLT is actually a mathematical statement about the standardized version $\bar{x}$, namely

$$Zn = \sqrt{n}\left(\frac{\overline{x} - \mu}{\sigma}\right)$$

The central limit theorem says that as n increases, the sampling distribution of Z_n may be more and more closely approximated by a standard normal distribution. In other words, when the sample size n is large, then probabilities like $P\sigma\left(Z_n \leq Z^1\right)$ are very close to the probability $P\sigma\left(Z \leq Z^1\right)$ where Z is a standard normal random variable.

Consider the events $Z_n \leq Z^1$

i.e. $\quad Z_n \frac{\sqrt{n}\left(\overline{x} - \mu\right)}{\sigma} \leq Z^1$

$$\overline{x} \leq \mu \neq Z^{11}\frac{\sigma}{\sqrt{n}}$$

Thus

$$P\sigma\left[Z_n \leq Z^1\right] = P\sigma\left[\frac{\sqrt{n}(\overline{\times} - \mu) \leq Z^1}{\sigma}\right]$$

$$= \qquad P\sigma\left(\overline{\times} \leq \mu + Z^1 \tfrac{\sigma}{\sqrt{n}}\right)$$

When n is large, CLT tells us that this probability is approximately the probability $P\sigma\left[Z \leq Z^1\right]$obatainable from the standard normal table. We can use the standard normal distribution to approximate probabilities whose exact values are obtainable only from the actual sampling distribution of $\overline{\times}$provided the sample size is large.

According to the CLT, we use the standard normal distribution to approximate the probability of events of the form

$$-Z^1 \leq \frac{\sqrt{n}(\overline{\times} - \mu) \leq Z^1}{\sigma}$$

Which can be written in the equivalent form as

$$-Z^1 \frac{\sigma}{\sqrt{n}} = \overline{\times} - \mu \leq Z^1 \frac{\sigma}{\sqrt{n}}$$

Using the normal approximation and expression. We set $L = Z^1\left(\sigma / \ \sqrt{n}\right)$.

Thus the probability of the event is

$$-L \leq \overline{\times} - \mu \leq L$$

Thus L is called the level of accuracy and is defined as

$$L = \frac{Z\sigma}{\sqrt{n}}$$

To find how large must the sample size n be, we algebraically solved from the L.

i.e $\qquad L = Z^1 \dfrac{\sigma}{\sqrt{n}}$

$$\sqrt{n}L = Z \ \sigma$$

$$\sqrt{n} = \frac{Z^1\sigma}{L}$$

$$n = \frac{Z^2\sigma^2}{L}$$

Here Z^1 will be determined by the preassigned value of P, that is the confidence level.

d) The Chi-square Distribution $\left(\times^2\right)$

Although the estimation of means is an important statistical problem variances are also important. Parameters, and their estimation can lead to the consideration of sampling distribution known as the chi-square distribution. Assume that $\times_1, \times_2, ----\times_n$ is a random sample from an $N(\mu, \sigma^2)$ population. Since we know that the sample mean

$$\overline{\times} = \frac{1}{n}\sum_{i=1}^{n} \times_i$$

has an $\mathrm{N}\left(\mu, \sigma^2/n\right)$ distribution. Now consider the distribution of the sample variance.

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\times_i - \overline{\times}\right)^2$$

rather than study he distribution of S^2 directly, it is customary to study the distribution of

$$\frac{nS^2}{\sigma^2} = \sum_{i=1}^{n}\frac{\left(\times_i - \overline{\times}\right)^2}{\sigma^2}$$

The distribution of this random variable depends on a single parameter, called degrees of freedom and in this case the degree of freedom parameter has value n-1.

Thus the degree of freedom (d.f) of nS^2/σ^2 is completely known once the value of n is known regardless of whether or not the value of σ^2 is known. It is this property that makes the distribution of nS^2/σ^2 useful for statistical inference. The distribution of nS^2/σ^2 is called chi-square with n-1 degrees of freedom and is abbreviated as $\times^2 (n-1)$.

e. The student t distribution

Another distribution of great importance in many statistical applications is the t distribution. The distribution is useful for researchers particularly involved with the estimation of means especially when the sample size n is small.

Like the chi-square distribution, the t distribution depends on a single parameter called degrees of freedom, which is denoted are v. The t distribution is symmetric about the origin, but it has fatter tail than the normal distribution. The t distribution approaches the normal distribution as the degrees of freedom increases. If Z is $\mathrm{N}(0,1))$ and Let Y be a chi square $\times^2 (V)$ random variable that is independent of Z. Then the $\sigma.v$ defined by

$$t = \frac{\mathrm{Z}}{\sqrt{\mathrm{Y}/v}}$$

has a $t(v)$ distribution.

Thus we can form a t distribution with v.d.f. by dividing a $\mathrm{N}(0,1))$ $\sigma.v$ by the square root of an independent chi-squared $\sigma.v$ divided by its d.f.

Let $X_1, X_2, ---- X_n$ denote a random sample from $N(\mu, \sigma^2)$ population, and Let $\overline{\times}$ and S^2 be the sample mean and variance. It is possible to show that the $\sigma.v's$ Z and Y are independent since Z = $Z = \sqrt{n}(\overline{\times}) - \mu \backslash \sigma$ has an $N(0,1))$ distribution and $Y^1 = ns^2 \backslash \sigma^2$ has a $\times^2 (n-1)$ distribution.

However it is clear that

$$t = \frac{Z}{\sqrt{Y/v}} = \frac{\left[\sqrt{n(\overline{\times}-\mu)}\right]}{\sigma} \Bigg/ \left(\frac{\sqrt{ns^2}}{\sigma^2} \backslash n-1\right)$$

$$= \frac{\sqrt{n}(\overline{\times}-\mu)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\times_i-\overline{\times})^2}}$$

This shows that the random variable and its distribution does not depend on σ^2. This means that the sample mean and the population mean may be related in a distribution that does not depend on σ^2. This fact allow us to make, inferences about the population, mean without knowing the population variance.

f. The F distribution

When applying the t-test to the means of samples, there was an assumption of a common parent variance. If the sample variances differ so much that we cannot assume a common parent variance, the t-test is nullified. The F-test is a variance ratio test. Suppose U and V are independent random variables such that U has a $\times^2 (V_1)$ distribution and V has a $\times^2 (V_2)$ distribution. The random variable defined by the ratio $F = \frac{U \backslash v_1}{V \backslash v_2}$

Is said to have an F-distribution with v_1 and v_2 d.f. This F-distribution is denoted by the symbol $F(v_1, v_2)$.

The distribution for F is not symmetrical and it is convenient in practice to calculate the ratio F with the larger variance in the numerator so that $F_{v_2}^{v_1}$ is always greater than one, and v_1 is the number of d.f. for the greater variance. This allows for a simpler tabulation of probabilities.

Statistical tests and Procedures

The goal of scientific research can be either descriptive/predictive or causal. Generally, the analytical emphasis is on obtaining results that can be generalized. To accomplish this data obtained is subjected to theoretical and philosophical rules of probability. The purpose of any analytical method (or statistical method) is to convert data into information needed to make decisions.

Data transformations are employed in statistical methods since the purpose of statistical methods is to extract information, in an understandable form from sets of observations. In statistics one usually work with data have been transformed in some way rather than using the original numbers. It is therefore worth mentioning the man data transformations available. The data transformations are:

i. rounding

ii. grouping

iii. dividing or multiplying by a constant

iv. differencing

v. taking logarithms

vi. taking the reciprocal

vii. deflating

it is essential to remember that statistical inferences are generalizations on the basis of limited information. To describe the distribution of any statistical measure whose value is estimated from a sample of data, the term sampling distribution is used. Note that a sample of data provides values which are manipulated according to a prescribed formula to obtain an estimate of the value of some desired feature.

Sampling is carried out to enable estimates to be made of characteristics of the population. It is essential to remember that the population mean μ and variance σ^2 are constants. These are called population parameters. Most often the mean and variance of the population are the characteristics whose values are to be estimated. By contrast, the sample mean $\bar{\times}$ and sample variance S^2 are random variables, varying from sample to sample, with a certain probability distribution. For example, the distribution of $\bar{\times}$ was found to be approximately normal. A random variable such as $\bar{\times}$ or S^2, which is calculated from the observations, in a sample, is given the technical name sample, statistic.

A statistic is a function of the sample items.

From a practical viewpoint, it is important to infer information about a population by the use of samples drawn from it. Such problems are discussed by the use of sampling theory. An important statistical problem is the choice of a sample statistic for making inference about the population parameter. These problem fall within the province of estimation theory.

Estimation of Parameters

To generalize, consider any population parameter θ, and denote its estimator by $\hat{\theta}$.

Definition: Estimator

An estimator g $(\hat{\theta})$ of the function g (θ) of a population parameter, θ is any statistic (known function of the random sample which is itself a random variable independent of any unknown parameters) whose values are used to estimate g (θ).

In other words, the estimator is the procedure used to obtain an estimate. We say that g $(\times_1, \times_2, ---- \times_n)$ estimates g (θ) to emphasize the dependence of the estimator on the random sample. The particular value g $(\times_1, \times_2, ---- \times_n)$ is an estimate of g (θ).

The desirable properties of estimators are

- Unbiasedness
- Minimum variance
- Consistency
- Relative efficiency
- Efficiency
- Sufficiency

Statistical estimators are referred to as point estimators when they are single numbers or points on the red axis used to estimate the population parameters to distinguish them from interval estimators which involve two points constructed from the point estimators which contain a given parameter with a specified probability or level of confidence.

Formally, to generate any population θ , and its estimator $\hat{\theta}$.

$\hat{\theta}$ is an unbiased estimator of θ iff

$$\mathrm{E}\left(\hat{\theta}\right) = \theta$$

An estimator $\hat{\theta}$ is called biased if $\mathrm{E}\left(\hat{\theta}\right)$ is different from $\mathrm{E}\left(\hat{\theta}\right)$. In fact, bias is defined as this difference.

$$\text{Bias} = \mathrm{E}\left(\hat{\theta}\right) - \theta.$$

The sample mean squared deviation defined as

$$\text{MSD} = \frac{1}{n}\sum_{i=1}^{n}(\times_i - \times_1)^2$$

is a biased estimator. It will on the average, under estimate the population variance σ^2.

If

$$S^2 \frac{1}{n-1}\sum_{i=1}^{n}\left(\times_i - \overline{\times}\right)^2$$

This is an unbiased estimator of σ^2. Both sample mean and median are unbiased estimators of $\sigma^2\mu$ in a normal population.

If the mean of a sampling distribution of a statistic equals the corresponding population parameter, the statistic is called an unbiased estimator of the parameter, otherwise it is called a biased estimator. The corresponding values of such statistics are called unbiased or biased estimator respectively.

If the sampling distributions of two statistics have the same mean (or expectation), the statistic with the smaller variance is called an efficient estimator of the mean while the other statistic is called an inefficient estimator. The corresponding values of the statistics are called efficient or inefficient estimates respectively.

If we consider all possible statistics whose sampling distributions have the same mean, the one with the smallest variance is sometimes called the most efficient or best estimator of this mean. We describe $\hat{\theta}$ as more efficient because it has smaller variance.

The relative efficiency of two estimators is defined as for unbiased estimators, the relative efficiency of $\hat{\theta}$ compared to

$$\hat{\theta} = \frac{Var\hat{\theta}}{\operatorname{var}\hat{\theta}}$$

An estimator that is more efficient than any other is called absolutely efficient, or simply efficient.