MODULE 7 PART 3

Population and Sample-sampling and sample designs

In statistics, we are interested in obtaining information about a total collection of elements,which we will refer to as the population. The population is often too large for us to examine each of its members. For instance, we might have all the residents of a given state, or all the television sets produced in the last year by a particular manufacturer, or all the households in a given community. In such cases, we try to learn about the population by choosing and then examining a subgroup of its elements. This subgroup of a population is called a **sample.**

If the sample is to be informative about the total population, it must be, in some sense,representative of that population. For instance, suppose that we are interested in learning about the age distribution of people residing in a given city, and we obtain the ages of the first 100 people to enter the town library. If the average age of these 100 people is 46.2years, are we justified in concluding that this is approximately the average age of the entire population?

Probably not, for we could certainly argue that the sample chosen in this case is probably not representative of the total population because usually more young students and senior citizens use the library than do working-age citizens. In certain situations, such as the library illustration, we are presented with a sample and must then decide whether this sample is reasonably representative of the entire population.

In practice, a given sample generally cannot be assumed to be representative of a population unless that sample has been chosen in a random manner. This is because any specific nonrandom rule for selecting a sample often results in one that is inherently biased toward some data values as opposed to others.

Thus, although it may seem paradoxical, we are most likely to obtain a representative sample by choosing its members in a totally random fashion without any prior considerations of the elements that will be chosen. In other words, we need not attempt to deliberately choose the sample so that it contains, for instance, the same gender percentage in the same percentage of people in each profession as found in the general population.

Rather, we should just leave it up to "chance" to obtain roughly the correct percentages.

Once a random sample is chosen, we can use statistical inference to draw conclusions about the entire population by studying the elements of the sample

Sampling

Sampling is a statistical procedure that is concerned with the selection of certain individual observations from the target population. It helps to make statistical inferences about the population. Some of the basic terminologies are as follows

Population

A population is any complete group (i.e., people, sales territories, stores, etc.) sharing some common set of characteristics. It can be defined as including all people or items with the characteristic one wishes to understand and draw inferences about them.

Population frame

A list, map, directory, or other source used to represent the population

Census

A census is an investigation of all the individual elements making up the population—a total
listing rather than a sample.

Sample

A sample is a subset or some part of a larger population. It is "a smaller (but hopefully

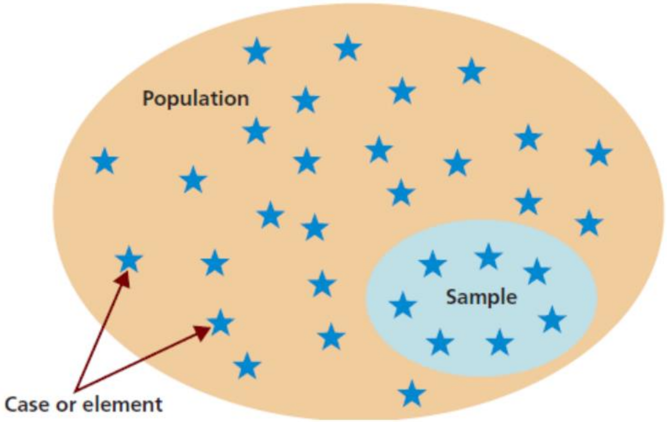
representative) collection of units from a population used to determine truths about that
population”

SAMPLE DESIGN

A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample.

Steps in Sample Design:

- **Type of Universe**
- **Sampling Unit**
- **Sampling Frame**
- **Size of Sample**
- **Budgetary Constraints**
- **Sampling Procedure**



Diagrammatic representation of sampling process

Characteristics Of A Good Sample Design

- Sample design must result in a truly representative sample
- Sample design must be such which results in a small sampling error
- Sample design must be viable in the context of funds available for the

- research study
- Sample design must be such so that systematic bias can be controlled in a better way
- Sample should be such that the results of the sample study can be applied, in general, for the universe with a reasonable level of confidence

Basic principles of sampling

Theory of sampling is based on the following laws

a. Law of Statistical Regularity – This law comes from the mathematical theory of probability. According to King," Law of Statistical Regularity says that a moderately large number of the items chosen at random from the large group are almost sure on the average to possess the features of the large group." According to this law the units of the sample must be selected at random.

b. Law of Inertia of Large Numbers – This law states that the other things being equal the larger the size of the sample; the more accurate the results are likely to be.

Types of Sample Design

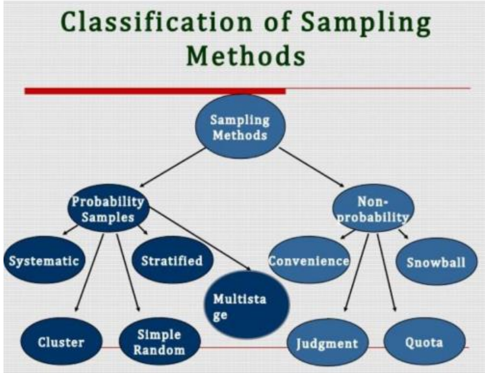
• **Probability Sampling Design**

Each element/respondent has a known probability of being included in the sample.

• **Non-probability Sampling Design**

Each element/respondent in the population is not given an equal chance of selection.

Probability Sampling Design



Types of Probability Sampling

1. Simple Random Sampling

2. Systematic Sampling

3. Stratified Random Sampling

a. Proportionate

b. Disproportionate

4. Cluster (or Area) Sampling

5. Multistage sampling

Types of Random Sampling

1. Simple Random Sampling

It is a sampling procedure where each element in the population will have an equal chance of being selected in the sample. This process is simple because it requires only one stage of the sample selection process. Here we number each frame unit from 1 to N. Then use a random number table or a random number generator to select n distinct numbers between 1 and N, inclusively. It is easier to perform for small populations but cumbersome for large populations.

2.Systematic Random Sampling

It is convenient and relatively easy to measure. Here an initial starting point is selected by a random process; then every nth number on the list is selected. The first sample element is selected randomly from the first k population elements. Thereafter, sample elements are selected at a constant interval, k from the ordered sequence frame.

k = N/n

where:

n= sample size

N= population size

k = size of selection interval

For example one wishes to take a sample of 50 from a list consisting of 10,000 purchase

orders. Purchase orders for the previous fiscal year are serialized 1 to 10,000 (N =10,000). A sample of fifty (n = 50) purchase orders is needed for an audit. k = 10,000/50= 200. First sample element randomly selected from the first 200 purchase orders.

Assume the 45th purchase order was selected. Subsequent sample elements: 245, 445,645 . . .

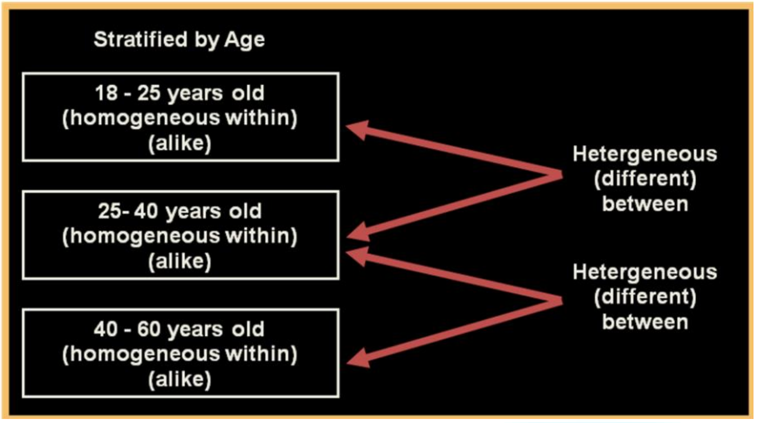
Stratified Random Sampling

Here the population is divided into non overlapping subpopulations called strata. A random sample is selected from each stratum. Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.

Every unit in a stratum has the same chance of being selected.

a) Proportionate -- the percentage of the sample taken from each stratum is proportionate to the percentage that each stratum is within the population.

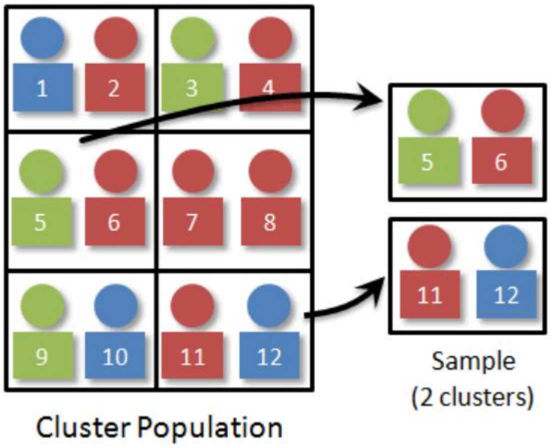
b) Disproportionate -- proportions of the strata within the sample are different from the proportions of the strata within the population.

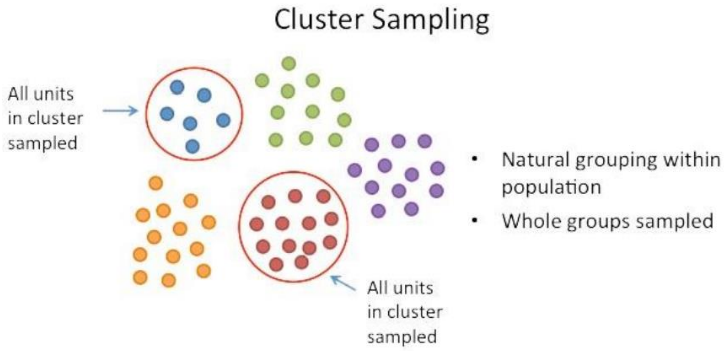

Stratified by Age
18 - 25 years old (homogeneous within) (alike)
25- 40 years old (homogeneous within) (alike)
40 - 60 years old (homogeneous within) (alike)
Hetergeneous (different) between
Hetergeneous (different) between

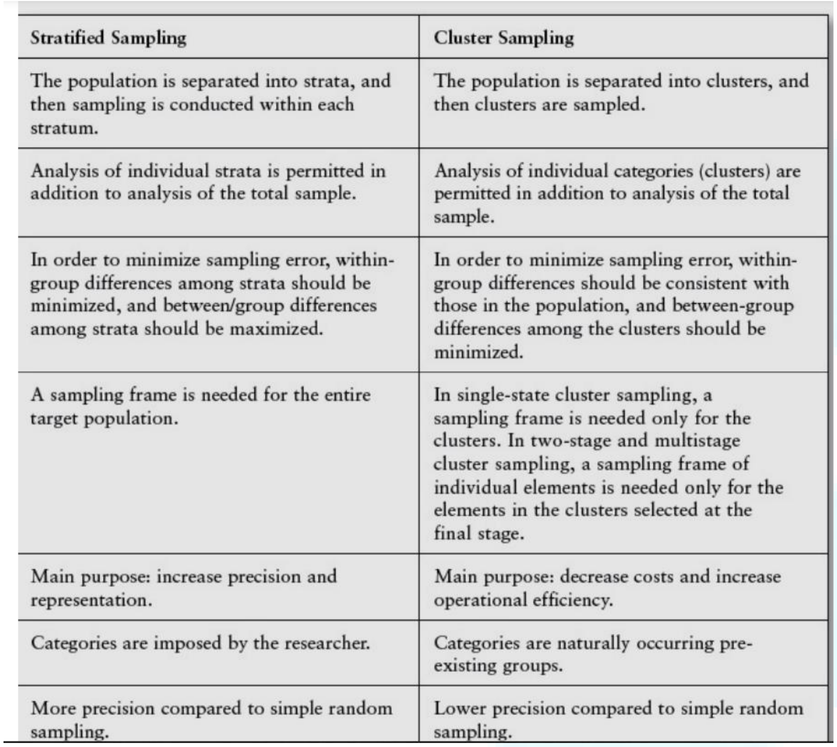
Cluster Sampling

It is also called 'two-stage sampling'. In the first stage a sample of areas is chosen. In the second stage a sample of respondents within those areas is selected. Here population is divided into non overlapping clusters or areas of homogeneous units usually based on

geographical dispersed population. Each cluster is a miniature, or microcosm, of the population. A subset of the clusters is selected randomly for the sample. If the number of elements in the subset of clusters is larger than the desired value of n, these clusters may be subdivided to form a new set of clusters and subjected to a random selection process.




Cluster Sampling
All units in cluster sampled
Natural grouping within population
Whole groups sampled
All units in cluster sampled



Stratified Sampling	Cluster Sampling
The population is separated into strata, and then sampling is conducted within each stratum.	The population is separated into clusters, and then clusters are sampled.
Analysis of individual strata is permitted in addition to analysis of the total sample.	Analysis of individual categories (clusters) are permitted in addition to analysis of the total sample.
In order to minimize sampling error, within-group differences among strata should be minimized, and between/group differences among strata should be maximized.	In order to minimize sampling error, within-group differences should be consistent with those in the population, and between-group differences among the clusters should be minimized.
A sampling frame is needed for the entire target population.	In single-state cluster sampling, a sampling frame is needed only for the clusters. In two-stage and multistage cluster sampling, a sampling frame of individual elements is needed only for the elements in the clusters selected at the final stage.
Main purpose: increase precision and representation.	Main purpose: decrease costs and increase operational efficiency.
Categories are imposed by the researcher.	Categories are naturally occurring pre-existing groups.
More precision compared to simple random sampling.	Lower precision compared to simple random sampling.

Multistage Sampling

Multi-stage sampling (also known as multi-stage cluster sampling) is a more complex form of cluster sampling which contains two or more stages in sample selection.

In multi-stage sampling large clusters of population are divided into smaller clusters in several stages in order to make primary data collection more manageable in terms of cost effectiveness andtime effectiveness.

It is quite effective in primary data collection from geographically dispersed population where face-to-face contact is required (e.g. semi-structured in-depth interviews).

Types of Non Probability Sampling

1) Convenience Sampling:

A type of nonprobability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, readily available and convenient. It is also termed as grab or opportunity sampling or accidental or haphazard sampling. Sample elements are selected for the convenience of the researcher. The researcher using such a sample cannot scientifically make generalizations about the total population
from this sample because it would not be representative enough. This type of sampling is most useful for pilot testing.

2) Judgment Sampling: Here the sample elements are selected by the judgment of the researcher. The researcher chooses the sample based on who they think would be appropriate
for the study. This is used primarily when there are a limited number of people that have
expertise in the area being researched.

3) Quota Sampling: Here the population is first segmented into mutually exclusive subgroups, just as in stratified sampling. Then judgment is used to select subjects or units from
each segment based on a specified proportion. In quota sampling the selection of the sample
is non-random. For example, an interviewer may be told to sample 200 females and 300males between the ages of 45 and 60. He might be tempted to interview those who look most helpful.

4.Snowball Sampling: survey subjects are selected based on referral from other survey respondents. In social science research, snowball sampling is a similar technique, where existing study subjects are used to recruit more subjects into the sample.

Types of Error

i)Sampling error

If researchers are not careful in planning and defining the sampling process, it can lead to faulty research findings. Sampling error is the error that occurs because of a representative sample from the population rather than the entire population.

In statistical terminology, it's the difference between the statistic you measure and the parameter you would find if you took a census of the entire population. Sample error can't be eliminated, but it can be reduced. In general, it works like the larger the sample, the smaller the margin of error.

ii) Non Sampling error

This is due to poor data collection methods (like faulty instruments or inaccurate data recording, missing data, selection bias, non response bias (where individuals don't want to or can't respond to a survey), poorly conceived concepts, vague definitions and defective questions. Increasing the sample size will not reduce these errors. They key is to avoid making the errors in the first place with a well-planned design for the survey or experiment.

Parameter

In a statistical inquiry, our interest lies in one or more characteristics of the population. A measure of such a characteristic is called a parameter. For example, we may be interested in the mean income of the people of some region for a particular year. We may also like to know the standard deviation of these incomes of the people. Here, both mean and standard deviation are parameters.

Parameters are conventionally denoted by Greek alphabets. For example, the population mean can be denoted by p and population standard deviation can be i denoted by o .

It is important to note that the value of a parameter is computed from all the population observations. Thus, the parameter 'mean income' is calculated from all the income figures of different individuals that constitute the population.

Similarly, for the calculation of the parameter 'correlation coefficient of heights and weights', we require the values of all the pairs of heights and weights in a population. Thus,

We can define a parameter as a function of the population values. If 8 is a parameter that we want to obtain from the population values XI, X, , . X, , then

Statistic

While discussing the census and the sample survey, we have seen that due to various constraints, sometimes it is difficult to obtain information about the whole population. In other words, it may not always be possible to compute a population parameter. In such situations, we try to get some idea about the parameter from the information obtained from a sample drawn from the population.

This sample information is summarized in the form of a stati.vtic. For example, sample mean or sample median or sample mode is called a statistic. Thus, a statistic is calculated from the values of the units that are included in the sample. So, a statistic can be defined as the function of the sample values. Conventionally, a statistic is denoted by an English alphabet.

For example, the sample mean may be denoted by 2 and the sample standard deviation may be denoted by s. If T is a statistic that we want to obtain from the sample values x, , x, , . xn , then

Estimator and Estimate

The basic purpose of a statistic is to estimate some population parameter. The Procedure followed or the formula used to compute a statistic is called an estimator and the value of a statistic so computed is known as an estimate. .

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This formula is an estimator. Next, if we use this formula and get jj = 10, this ' 10' is an estimate.

SAMPLING DISTRIBUTION

By now it should be clear that generally the size of a sample is much smaller than the parent population. Consequently, many samples can be selected from the same population which are different from one another. Since an estimate of a parameter depends upon the sample values, and these values may change from one sample to another, there can be different estimates or values of a statistic for the same parameter.

This variation in values is called sampling fluctuation.

Suppose, a number of samples, each of size n, are drawn from a population of size N and for each sample, the value of the statistic is computed. If the number of samples is large, these values can be damaged in the form of a relative frequency distribution.

When the number of samples tends to infinity, the resultant relative frequency distribution of the values of a statistic is called the sampling distribution of the given statistics.

Suppose, we are interested in estimating the population mean (which is a parameter), denoted by p. A random sample of size n is drawn from this population (of size N). The sample mean

$$\bar{x} = \frac{1}{n}\sum x_i$$

is a statistic corresponding to the population mean p. We should note x that is a random variable as its value changes from one sample 'to another in a probabilistic manner

STANDARD ERROR OF A STATISTIC

In the previous Section we learnt that we can draw a number of samples depending upon the population and sample sizes. From each sample we get a different value for the statistic we qe looking for. These values can be arranged in the form of a probability distribution, which is called the sampling distribution of the concerned statistic.

The statistic is also similar to a random variable since a probability is

' attached to each value it takes.

two important properties of the sampling distribution.

1) The expectation of the sampling distribution of the statistic is equal to the population parameter. Thus if we have the sampling distribution of sample means, then its expected value is equal to population mean. Symbolically,

$$E(\bar{x}) = \mu .$$

2) The standard deviation of the sampling distribution is called 'standard mr' of the concerned statistic. Thus if we have sampling distribution of sample means, then its standard deviation is called the 'standard error of sample means'. Thus stand nmr indicates the spread of the sample means away film the population mean.

Measurement and Scaling techniques

scaling techniques

Scaling technique is a method of placing respondents in continuation of gradual change in the pre-assigned values, symbols or numbers based on the features of a particular object as per the defined rules. All the scaling techniques are based on four pillars, i.e., order, description, distance and origin.

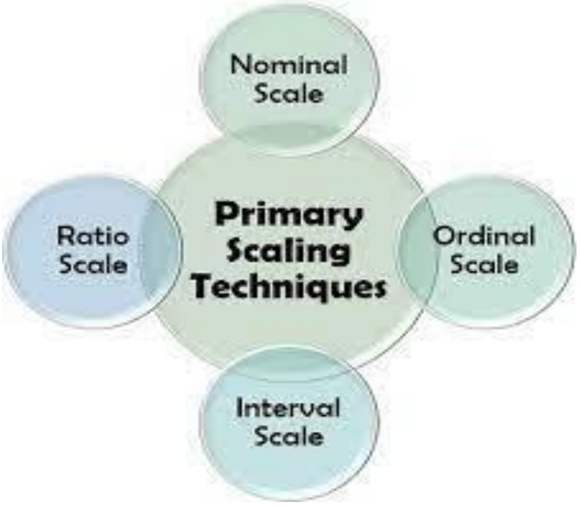
Types of Scaling Techniques

The researchers have identified many scaling techniques; today, we will discuss some of the most common scales used by business organizations, researchers, economists, experts, etc.

These techniques can be classified as primary scaling techniques and other scaling techniques.

Primary Scaling Techniques

The major four scales used in statistics for market research consist of the following:



Nominal Scale

Nominal scales are adopted for non-quantitative labeling variables which are unique and different from one another.

Types of Nominal Scales:

- Dichotomous: A nominal scale that has only two labels is called ‘dichotomous’; for example,Yes/No.
- Nominal with Order: The labels on a nominal scale arranged in an ascending or descending order are termed as ‘nominal with order’; for example, Excellent, Good, Average, Poor, Worst.
- Nominal without Order: Such nominal scale which has no sequence, is called ‘nominal without order’; for example, Black, White.

Ordinal Scale

The ordinal scale functions on the concept of the relative position of the objects or labels based on the individual’s choice or preference.

For example, At Amazon.in, every product has a customer review section where the buyers rate the listed product according to their buying experience, product features, quality, usage,etc.

The ratings so provided are as follows:

1. 5 Star – Excellent
2. 4 Star – Good
3. 3 Star – Average
4. 2 Star – Poor
5. 1 Star – Worst

Interval Scale

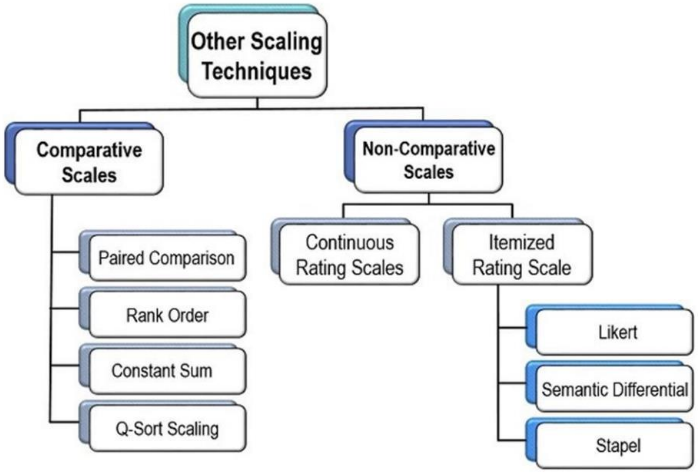
An interval scale is also called a cardinal scale which is the numerical labeling with the same difference among the consecutive measurement units. With the help of this scaling technique,
researchers can obtain a better comparison between the objects.

Ratio Scale

One of the most superior measurement techniques is the ratio scale. Similar to an interval scale, a ratio scale is an abstract number system. It allows measurement at proper intervals,
order, categorization and distance, with an added property of originating from a fixed zero point. Here, the comparison can be made in terms of the acquired ratio.

Other Scaling Techniques

Scaling of objects can be used for a comparative study between more than one object(products, services, brands, events, etc.). Or can be individually carried out to understand the
consumer's behavior and response towards a particular object

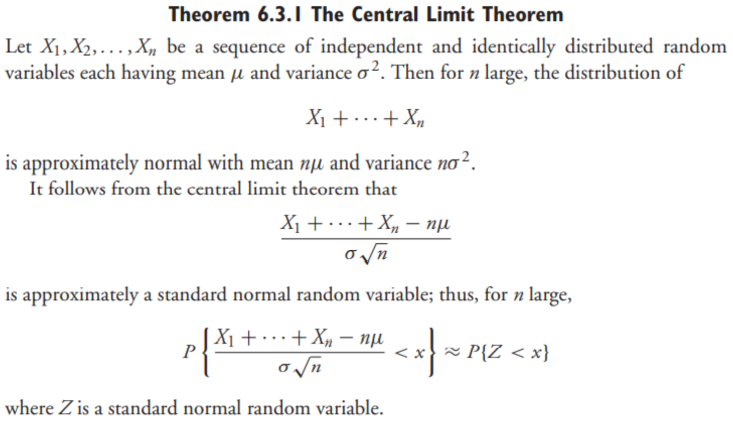


THE CENTRAL LIMIT THEOREM

In this section, we will consider one of the most remarkable results in probability —namely, the central limit theorem. Loosely speaking, this theorem asserts that the sum of a large number of independent random variables has a distribution that is approximately normal.

Hence, it not only provides a simple method for computing approximate probabilities for sums of independent random variables, but it also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit a bell-shaped(that is, a normal) curve.

In its simplest form, the central limit theorem is as follows



Theorem 6.3.1 The Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for n large, the distribution of

$$X_1 + \cdots + X_n$$

is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

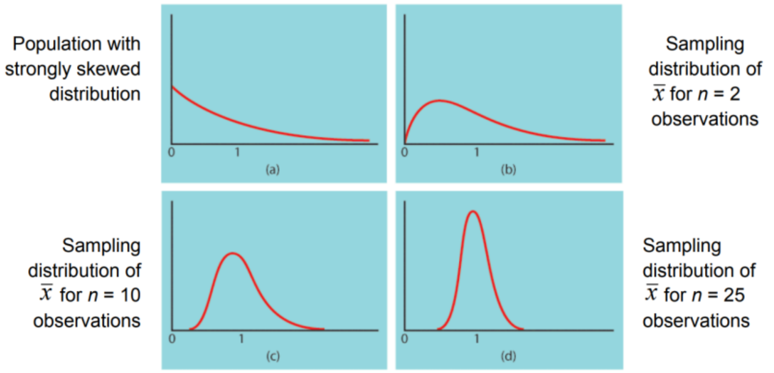
It follows from the central limit theorem that

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

is approximately a standard normal random variable; thus, for n large,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < x\right\} \approx P\{Z < x\}$$

where Z is a standard normal random variable.



The sampling distribution

The sampling distribution of a statistic is the distribution of all possible values taken by the statistic when all possible samples of a fixed size n are taken from the population. It is a theoretical idea—we do not actually build it.

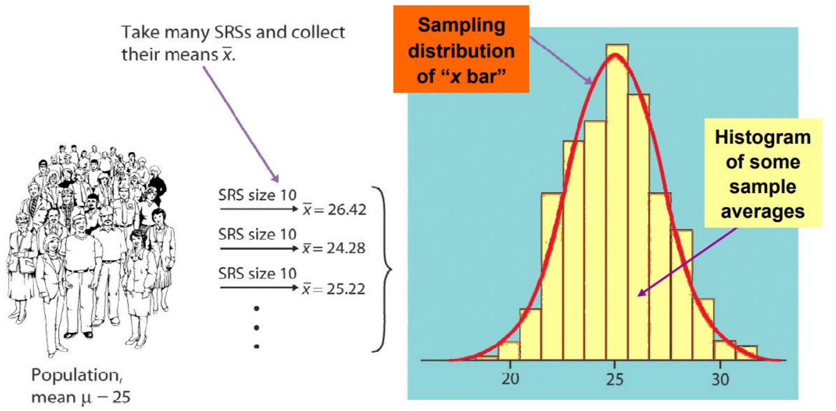
The sampling distribution of a statistic is the probability distribution of that statistic.

We take many random samples of a given size

n from a population with meanμ and standard deviation σ.

Some sample means will be above the population mean

μ and some will be below, making up the sampling distribution.


Take many SRSs and collect their means $\bar{x}$.
SRS size 10
$\bar{x}$ = 26.42
SRS size 10
$\bar{x}$ = 24.28
SRS size 10
$\bar{x}$ = 25.22
Population, mean μ − 25
Sampling distribution of "x bar"
Histogram of some sample averages
20
25
30

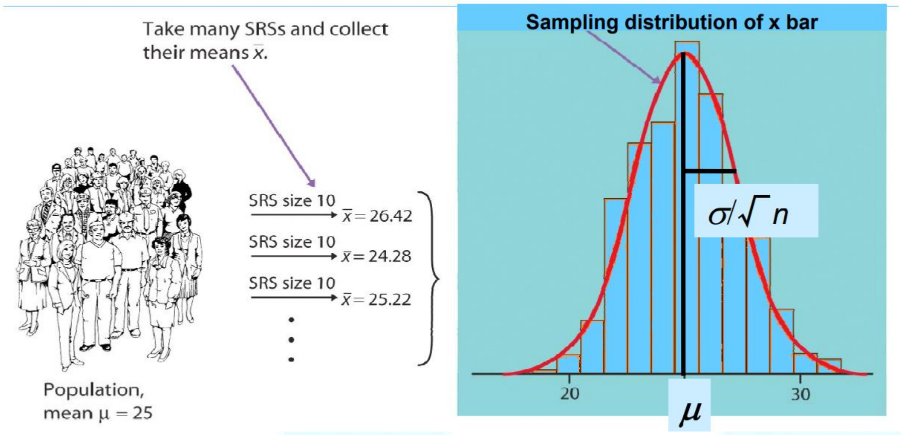
For any population with mean μ and standard deviation σ:

- The mean, or center of the sampling distribution of , is equal to the

population mean μ : $\mu : \; \mu_{\bar{x}} = \mu$.

- The standard deviation of the sampling distribution is σ/√n, where

n is the sample size : . $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.



Mean of a sampling distribution of

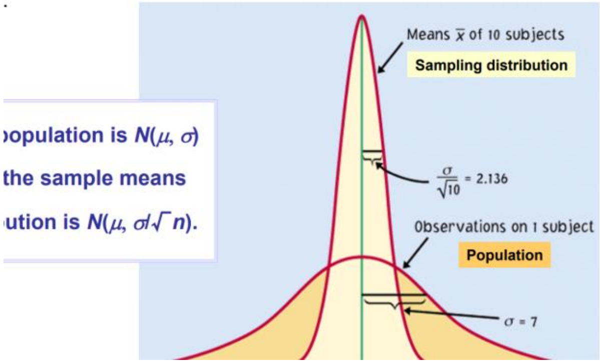
There is no tendency for a sample mean to fall systematically above or below μ, even if the distribution of the raw data is skewed. Thus, the mean of the sampling distribution is an unbiased estimate of the population mean μ — it will be "correct on average" in many samples.

Standard deviation of a sampling distribution of

The standard deviation of the sampling distribution measures how much the Sample statistics vary from sample to sample. It is smaller than the standard deviation of the population by a factor of √n. Î Averages are less variable than individual observations.

For normally distributed populations

When a variable in a population is normally distributed, the sampling distribution of for all possible samples of size n is also normally distributed.



F-DISTRIBUTION

As we have said in the previous unit, F-distribution was introduced by Prof. R. A. Fisher and defined as the ratio of two independent chi-square variates when divided by their respective degrees of freedom. If we draw a

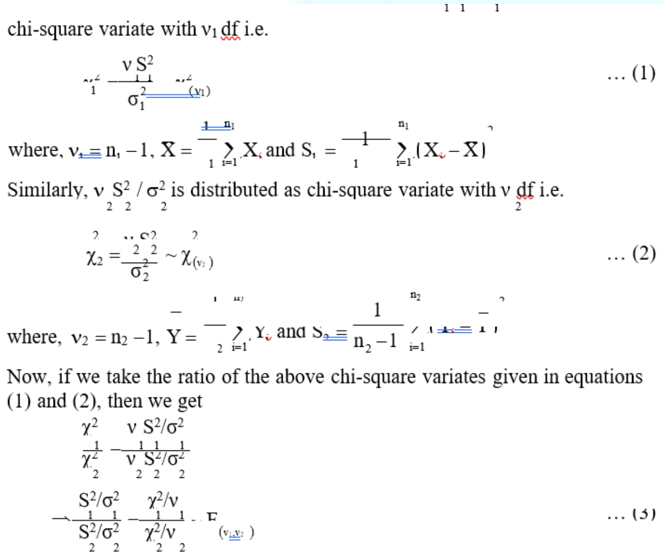
random sample X1, X2 ,…, Xn of size n1 from a normal population with mean

1 and variance σ2 and another independent random sample Y , Y ,…, Y of

1 1 2 n2

size n2 from another normal population with mean 2 and variance 2

respectively then as we have studied in Unit 3 that S2 / 2 is distributed as



chi-square variate with ν_1 df i.e.

$$\chi_1^2 = \frac{\nu_1 S_1^2}{\sigma_1^2} \sim \chi^2_{(\nu_1)} \qquad \ldots (1)$$

where, $\nu_1 = n_1 - 1$, $\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ and $S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2$

Similarly, $\nu_2 S_2^2 / \sigma_2^2$ is distributed as chi-square variate with ν_2 df i.e.

$$\chi_2^2 = \frac{\nu_2 S_2^2}{\sigma_2^2} \sim \chi^2_{(\nu_2)} \qquad \ldots (2)$$

where, $\nu_2 = n_2 - 1$, $\bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ and $S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$

Now, if we take the ratio of the above chi-square variates given in equations (1) and (2), then we get

$$\frac{\chi_1^2}{\chi_2^2} = \frac{\nu_1 S_1^2/\sigma_1^2}{\nu_2 S_2^2/\sigma_2^2}$$

$$\Rightarrow \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} \sim F_{(\nu_1, \nu_2)} \qquad \ldots (3)$$

PROPERTIES OF F-DISTRIBUTION

The F-distribution has wide properties in Statistics. Some of them are as follow:

1.The probability curve of F-distribution is a positively skewed curve. The curve becomes highly positively skewed when ν2 is smaller than ν1.

2.F-distribution curve extends on abscissa from 0 to .

3.F-distribution is a unimodal distribution, that is, it has a single mode.

4.The square of t-variate with ν df follows F-distribution with 1 and ν degrees of freedom.

degrees of freedom.

5. The mean of F-distribution with (ν_1, ν_2) df is $\frac{\nu_2}{\nu_2 - 2}$ for $\nu_2 > 2$.

6. The variance of F-distribution with (ν_1, ν_2) df is

$$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \text{ for } \nu_2 > 4.$$

7. If we interchange the degrees of freedom ν_1 and ν_2 then there exists a very useful relation as

$$F_{(\nu_1,\nu_2),(1-\alpha)} = \frac{1}{F_{(\nu_2,\nu_1),\alpha}}$$

Probability Curve of F-distribution

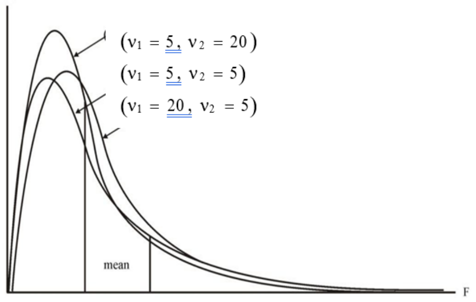


Fig. 4.1: Probability curves of F-distribution for (5, 5), (5, 20) and (20, 5) degrees of freedom.

The F-distribution has the following applications:

- F-distribution is used to test the hypothesis about equality of the variances of two normal populations.
- F-distribution is used to test the hypothesis about multiple correlation coefficients.

- F-distribution is used to test the hypothesis about correlation ratio.
- F-distribution is used to test the equality of means of k-populations, when one characteristic of the population is considered i.e. F-distribution is used in one-way analysis of variance.
- F-distribution is used to test the equality of k-population means for two characteristics at a time i.e. F-distribution is used in two-way analysis of variance.

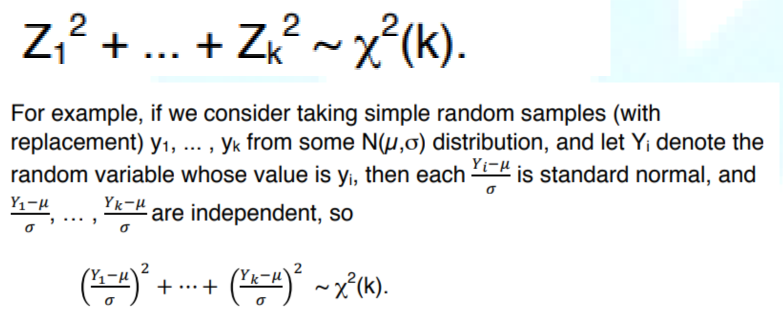
Chi-squared Distributions

Definition: The chi-squared distribution with k degrees of freedom is the distribution of a random variable that is the sum of the squares of k independent

standard normal random variables. We'll call this distribution χ2

(k).

Thus, if Z1, ... , Zk are all standard normal random variables (i.e., each Zi ~ N(0,1)), and if they are independent, then



$$Z_1^2 + \ldots + Z_k^2 \sim \chi^2(k).$$

For example, if we consider taking simple random samples (with replacement) $y_1, \ldots, y_k$ from some $N(\mu,\sigma)$ distribution, and let Y_i denote the random variable whose value is y_i, then each $\frac{Y_i-\mu}{\sigma}$ is standard normal, and $\frac{Y_1-\mu}{\sigma}, \ldots, \frac{Y_k-\mu}{\sigma}$ are independent, so

$$\left(\frac{Y_1-\mu}{\sigma}\right)^2 + \cdots + \left(\frac{Y_k-\mu}{\sigma}\right)^2 \sim \chi^2(k).$$

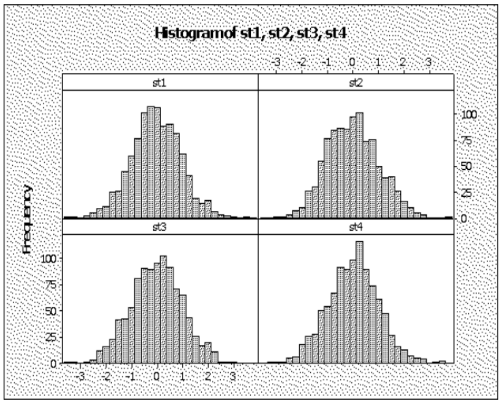
Notice that the phrase "degrees of freedom" refers to the number of independent standard normal variables involved. The idea is that since these k variables are independent, we can choose them "freely" (i.e., independently).

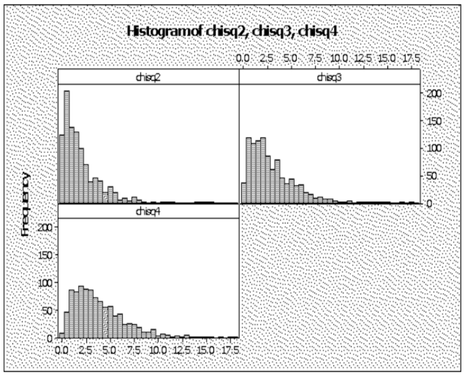
The following exercise should help you assimilate the definition of chi-squared distribution, as well as get a feel for the χ2(1) distribution.

For k > 1, it's harder to figure out what the χ2

(k) distribution looks like just using

the definition, but simulations using the definition can help. The following diagram
shows histograms of four random samples of size 1000 from an N(0,1) distribution:





$$
\begin{aligned}
f_Y(y) &= \frac{y^{(\nu/2-1)}e^{-y}}{\Gamma(\nu/2)}, \text{if } y > 0 \\
&= 0, \text{if } y \leq 0
\end{aligned}
$$

where $\Gamma(x)$ is the gamma function.

Note that this family of distributions is parametrised by ν, called the degrees of freedom. The graph of the density function for some members of this family ($\nu = 1, 3, 8, 10$) are shown in Fig.7 below

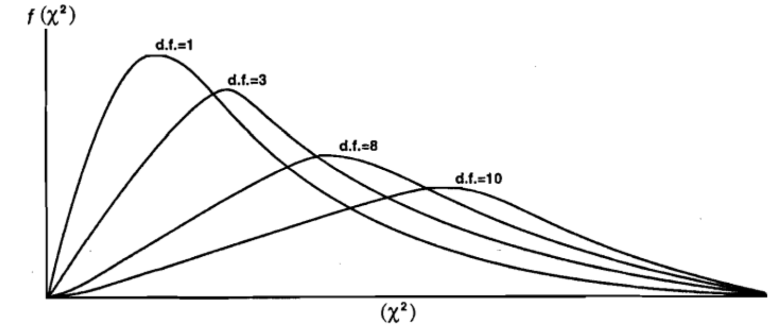


Fig.7 Various Chi-square distributions

t Distributions

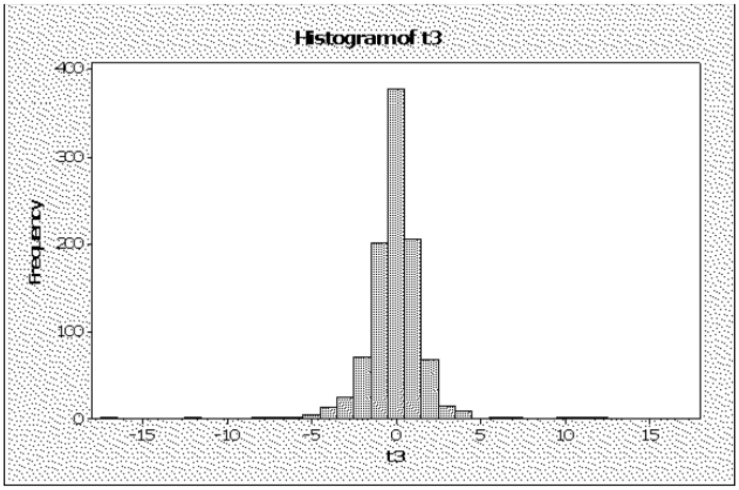
Definition: The t distribution with k degrees of freedom is the distribution of a random variable which is of the form

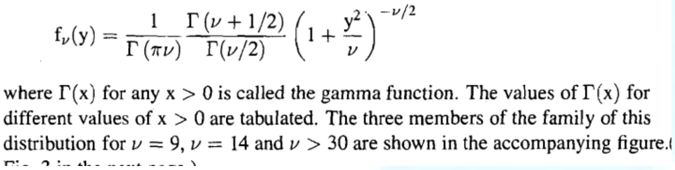
$$\frac{Z}{\sqrt{U/k}}$$

i. Z ~ N(0,1)

ii. U ~ χ2 (k), and

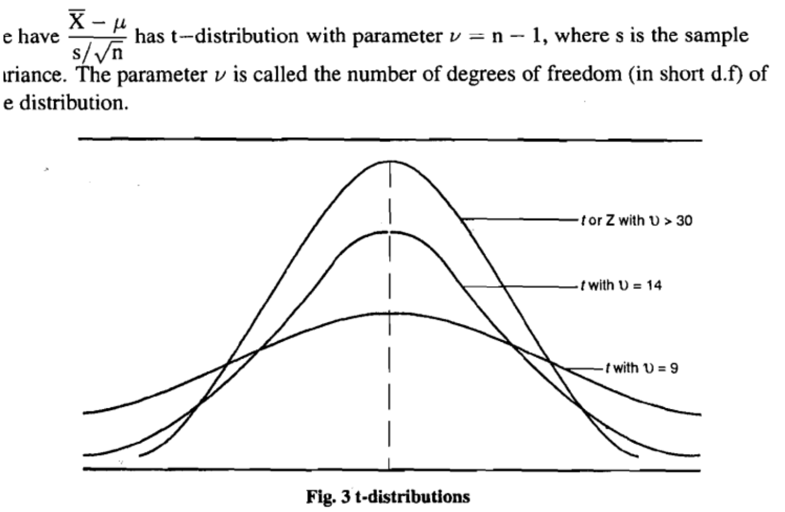
iii. Z and U are independent.





$$f_\nu(y) = \frac{1}{\Gamma(\pi\nu)} \frac{\Gamma(\nu + 1/2)}{\Gamma(\nu/2)} \left(1 + \frac{y^2}{\nu}\right)^{-\nu/2}$$

where $\Gamma(x)$ for any $x > 0$ is called the gamma function. The values of $\Gamma(x)$ for different values of $x > 0$ are tabulated. The three members of the family of this distribution for $\nu = 9$, $\nu = 14$ and $\nu > 30$ are shown in the accompanying figure.(Fig. 2 in the next page.)



e have $\frac{\overline{X}-\mu}{s/\sqrt{n}}$ has t−distribution with parameter $\nu = n - 1$, where s is the sample iriance. The parameter ν is called the number of degrees of freedom (in short d.f) of e distribution.

Fig. 3 t-distributions

THEORY OF ESTIMATION

In many real-life problems, the population parameter(s) is (are) unknown and someone is interested to obtain the value(s) of parameter(s). But, if the whole population is too large to study or the units of the population are destructive in nature or there is a limited resources and manpower available then it is not practically convenient to examine each and every unit of the population to find
the value(s) of parameter(s). In such situations, one can draw samples from the population under study and utilize sample observations to estimate the parameter(s).

Every one of us makes estimates(s) in our day to day life. For example, a house wife estimates the monthly expenditure on the basis of particular needs, a sweet shopkeeper estimates the sale of sweets on a day, etc. So the technique of finding an estimator to produce an estimate of the unknown parameter on the basis of a sample is called estimation.

There are two methods of estimation:

1. Point Estimation

Point Estimates

A point estimate is a sample statistic calculated using the sample data to estimate the most likely value of the corresponding unknown population parameter. In other words, point estimate is a single value derived from a sample and used to estimate the population value.

For instance, if we use a value of $\bar{x}$ to estimate the mean μ of a population.

$\bar{x} = \Sigma x/n$

2. Interval Estimation

A confidence interval estimate is a range of values constructed from sample data so that the population parameter is likely to occur within the range at a specified probability. The specified probability is the level of confidence.

Broader and probably more accurate than a point estimate

Used with inferential statistics to develop a confidence interval – where we believe with a certain degree of confidence that the population parameter lies.

Any parameter estimate that is based on a sample statistic has some amount of sampling error.

In statistics, interval estimation is the use of sample data to calculate an interval of possible values of an unknown population parameter.

PROPERTIES OF GOOD ESTIMATOR

Prof. Ronald A. Fisher was the man who pushed ahead the theory of estimation and introduced these concepts and gave some properties of good estimator as follows:

1. Unbiasedness

2. Consistency

3. Efficiency

4. Sufficiency

Unbiasedness

A first desirable property is that the expected value of the estimate ˆθ = T(y) be equal to the actual value of the parameter θ.

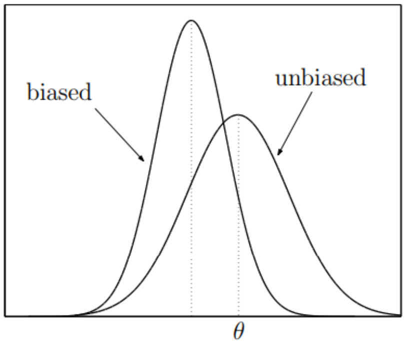
Definition 1.2. An estimator T(y) of the parameter θ is unbiased (or correct) In the above definition we used the notation

$$\mathbf{E}^{\theta}\left[T(\boldsymbol{y})\right] = \theta, \quad \forall\theta \in \Theta.$$

[·], which stresses the

dependency on θ of the expected value of T(y), due to the fact that the pdf of y is parameterized by θ itself.

The unbiasedness condition (1.1) guarantees that the estimator T(·) does not introduce systematic errors, i.e., errors that are not averaged out even when considering an infinite amount of observations of y. In other words, T(·) does not overestimate neither underestimate θ, on average



Consistency

Another desirable property of an estimator is to provide an estimate that converges to the actual value of θ as the number of measurements grows.

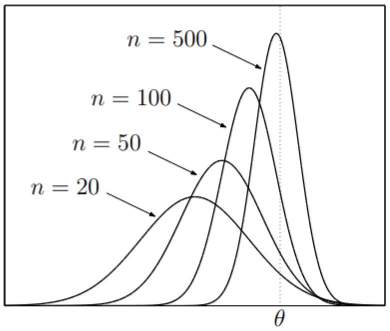
Being the estimate of a random variable, we need to introduce the notion of convergence in probability.

Definition 1.3. Let {yi}∞

i=1 be a sequence of random variables. The sequence of estimators ˆθn = Tn(y1 , . . . , yn) of θ is said to be consistent if ˆθn

converges in probability to θ, for all admissible values of θ, i.e.

$$\lim_{n\to\infty} \mathrm{P}\left(\left\|\hat{\theta}_n - \theta\right\| \geq \varepsilon\right) = 0, \quad \forall \varepsilon > 0, \quad \forall \theta \in \Theta.$$



ure 1.2: Probability density function of a consistent estimator.

In the previous section, we learned about unbiasedness. An estimator T is

said to be unbiased estimator of parameter, say, if the mean of sampling distribution of estimator T is equal to the true value of the parameter . This

The concept was defined for a fixed sample size. In this section, we will learn about consistency which is defined for increasing sample size.

If X ,X , ..., X 1 2 n is a random sample of size n taken from a population whose probability density (mass) function is f(x,θ) where, is the population

parameter then consider a sequence of estimators, say, T1 = t1(X1),

T2 = t2(X1, X2), T3 = t3(X1, X2, X3),..., Tn = tn(X1, X2, ..., Xn) . A sequence of estimators is said to be consistent for parameter if the deviation of the values

The estimator from the parameter tends to zero as the sample size increases. That means values of estimators tend to get closer to the parameter as sample size increases.

In other words, a sequence {Tn} of estimators is said to be consistent sequence of estimators of if Tn converges to in probability, that is

$$T_n \xrightarrow{p} \theta \text{ as } n \to \infty \quad \text{for every } \theta \in \Theta \qquad \ldots (3)$$

or for every $\varepsilon > 0$

$$\lim_{n \to \infty} P\left[|T_n - \theta| < \varepsilon\right] = 1 \qquad \ldots (4)$$

or for every$\varepsilon > 0$ and $\eta > 0$ there exist $n \geq m$ such that

$$P\left[|T_n - \theta| < \varepsilon\right] > 1 - \eta \; ; \qquad n \geq m \qquad \ldots (5)$$

where, m is some very large value of n. Expressions (3), (4) and (5) are to mean the same thing.

EFFICIENCY

In some situations, we see that there are more than one estimators of a parameter which are unbiased as well as consistent. For example, sample mean and sample median both are unbiased and consistent for the parameter

When sampling is done from normal population with mean and known variance σ2

.In such situations, there arises a necessity of some other criterion which will help us to choose 'best estimator' among them. A criterion which is based on the concept of variance of the sampling distribution of the estimator is termed as efficiency.

If T1 and T2 are two estimators of a parameter . Then T1 is said to be more efficient than T2 for all sample sizes if

$$\mathrm{Var}(T_1) < \mathrm{Var}(T_2) \qquad \text{for all } n$$

SUFFICIENCY

In statistical inference, the aim of the investigator or statistician may be to make a decision about the value of the unknown parameter (). The information that guides the investigator in making a decision is supplied by the random sample X ,X , ..., X 1 2 n . However, in most of the cases the observations would be too numerous and too complicated. Directly use of these observations

is complicated or cumbersome, therefore, a simplification or condensation would be desirable. The technique of condensing or reducing the random sample X ,X , ..., X 1 2 n into a statistic such that it contains all the information about parameters that is contained in the sample is known as sufficiency. So prior to continuing our search of finding the best estimator, we introduce the concept of sufficiency.

A sufficient statistic is a particular kind of statistic that condenses random samples X ,X , ...,X 1 2 n in a statistic T t(X ,X , ...,X) 1 2 n in such a way that no information about the parameter is lost. That means, it contains all the information about that contained in the sample and if we know the value of sufficient statistics, then the sample values themselves are not needed and cannothing tell you more about . In other words,

A statistic T is said to be sufficient for estimating a parameter if it contains all the information about which are available in the sample. This property of an estimator is called sufficiency. In other words,

An estimator T is sufficient for parameter if and only if the conditional distribution of X ,X , ...,X 1 2 n given T = t is independent of .

Mathematically,

f x ,x , ...,x / T t g x ,x , ...,x 1 2 n 1 2 n

where, the function g x , x ,...,x 1 2 n does not depend on the parameter .

Applications

Numerous fields require the use of estimation theory. Some of these fields include:

- **Interpretation of scientific experiments**

- **Signal processing**
- **Clinical trials**
- **Opinion polls**
- **Quality control**
- **Telecommunications**
- **Project management**
- **Software engineering**
- **Control theory (in particular Adaptive control)**
- **Network intrusion detection system**
- **Orbit determination**

Testing of hypothesis

The Null and Alternative Hypotheses

Rain from Seeded Clouds. In Example we modeled the log-rainfalls from 26 seeded clouds as normal random variables with unknown mean μ and unknown variance $\sigma2$. Let $\theta = (\mu, \sigma2)$ denote the parameter vector. We are interested in whether or not$\mu > 4$. To word this in terms of the parameter vector, we are interested in whether or not θ lies in the set $\{(\mu, \sigma2) : \mu > 4\}$. In Example 8.6.4, we calculated the probability that $\mu > 4$ as part of a Bayesian analysis. If one does not wish to do a Bayesian analysis, one must address the question of whether or not $\mu > 4$ by other means, such as those introduced in this chapter.

Consider a statistical problem involving a parameter θ whose value is unknown but must lie in a certain parameter space . Suppose now that can be partitioned into two disjoint subsets 0 and 1, and the statistician is interested in whether θ lies in 0 or in 1.

We shall letH0 denote the hypothesis that $\theta \in 0$ and letH1 denote the hypothesis that $\theta \in 1$. Since the subsets 0 and 1 are disjoint and 0 ∪ 1 = , exactly one of the hypotheses H0 and H1 must be true. The statistician must decide which of the hypotheses H0 or H1 appear to be true. A problem of this type, in which there are only two possible decisions, is called a problem of testing hypotheses. If the statistician makes the wrong decision, he might suffer a certain loss or pay a certain cost. In many problems, he will have an opportunity to observe some data before he has to make his decision, and the observed values will provide him with information about the value of θ. A procedure for deciding which hypothesis to choose is called a test procedure or simply a test.

Null and Alternative Hypotheses/Reject.

The hypothesis H0 is called the null hypothesis and the hypothesis H1 is called the alternative hypothesis. When performing a test, if we decide that θ lies in 1, we are said to reject H0. If we decide that θ lies in 0, we are said not to reject H0.

The terminology referring to the decisions in Definition 9.1.1 is asymmetric with regard to the null and alternative hypotheses. We shall return to this point later in the section.

H0: $\mu \geq 140$,

H1: $\mu < 140$.

The Critical Region and Test Statistics

Testing Hypotheses about the Mean of a Normal Distribution with Known Variance. Suppose that X = (X1,...,Xn) is a random sample from the normal distribution with unknown mean μ and known variance $\sigma 2$. We wish to test the hypotheses

H0: $\mu = \mu 0$,

H1: $\mu \quad = \mu 0$. (9.1.2)

It might seem reasonable to reject H0 if Xn is far from μ0. For example, we could choose a number c and reject H0 if the distance from Xn to μ0 is more than c. One way to express this is by dividing the set S of all possible data vectors x = (x1,...,xn) (the sample space) into the two sets

$S_0 = \{x : -c \le Xn - \mu 0 \le c\}$, and $S_1 = S_0^C$.

We then reject H0 if X $\in$ S1, and we don't reject H0 if X $\in$ S0. A simpler way to express the procedure is to define the statistic T = |Xn - μ0|, and reject H0 if T ≥ c.

In general, consider a problem in which we wish to test the following hypotheses:

H0: $\theta \in \Omega$, and H1: $\theta \in \Omega$. (9.1.3)

Suppose that before the statistician has to decide which hypothesis to choose, she can observe a random sample X = (X1,...,Xn) drawn from a distribution that involves the unknown parameter θ. We shall let S denote the sample space of then-dimensional random vector X. In other words, S is the set of all possible values of the random sample.

In a problem of this type, the statistician can specify a test procedure by partitioning the sample space S into two subsets. One subset S1 contains the values of X for which she will reject H0, and the other subset S0 contains the values of X for which she will not reject H0.

Test Statistic/Rejection Region. Let X be a random sample from a distribution that depends on a parameter θ. Let T = r(X) be a statistic, and let R be a subset of the real line. Suppose that a test procedure for the hypotheses (9.1.3) is of the form "reject

H0 if $T \in R$.” Then we call T a test statistic, and we call R the rejection region of the test.

When a test is defined in terms of a test statistic T and rejection region R, as in Definition 9.1.5, the set $S1 = \{x : r(x) \in R\}$ is the critical region from Definition 9.1.4.

Typically, the rejection region for a test based on a test statistic T will be some fixed interval or the outside of some fixed interval. For example, if the test rejects H0 when $T \geq c$, the rejection region is the interval $[c, \infty)$. Once a test statistic is being used, it is simpler to express everything in terms of the test statistic rather than try to compute the critical region from Definition 9.1.4. All of the tests in the rest of this book will be based on test statistics. Indeed, most of the tests can be written in the form “reject H0 if $T \geq c$.” (Example 9.1.7 is one of the rare exceptions.)

In Example 9.1.3, the test statistic is $T = |Xn - \mu 0|$, and the rejection region is the interval $[c, \infty)$. One can choose a test statistic using intuitive criteria, as in , or based on theoretical considerations. Some theoretical arguments are given in Sections 9.2–9.4 for choosing certain test statistics in a variety of problems involving a single parameter. Although these theoretical results provide optimal tests in the situations in which they apply, many practical problems do not satisfy the conditions required to apply these results.

Type I/II Error.

An erroneous decision to reject a true null hypothesis is a type I error, or an error of the first kind. An erroneous decision not to reject a false null hypothesis is called a type II error, or an error of the second kind.

In terms of the power function, if $\theta \in 0$, $\pi(\theta|\delta)$ is the probability that the statistician will make a type I error. Similarly, if $\theta \in 1$, $1 - \pi(\theta|\delta)$ is the probability of making a type II error. Of course, either $\theta \in 0$ or $\theta \in 1$, but not both.

Hence, only one type of error is possible conditional on θ, but we never know which it is.

If we have our choice between several tests, we would like to choose a test δ that has a small probability of error. That is, we would like the power function $\pi(\theta|\delta)$ to be low for values of $\theta \in 0$, and we would like $\pi(\theta|\delta)$ to be high for $\theta \in 1$. Generally, these two goals work against each other. That is, if we choose δ to make $\pi(\theta|\delta)$ small for $\theta \in 0$, we will usually find that $\pi(\theta|\delta)$ is small for $\theta \in 1$ as well.

For example,

the test procedure $\delta 0$ that never rejects H0, regardless of what data are observed, will have $\pi(\theta|\delta 0) = 0$ for all $\theta \in \Omega_0$. However, for this procedure $\pi(\theta|\delta 0) = 0$ for all

$\theta \in \Omega_1$ as well. Similarly, the test $\delta 1$ that always rejects H0 will have $\pi(\theta|\delta 1) = 1$ for all

$\theta \in \Omega_1$, but it will also have $\pi(\theta|\delta 1) = 1$ for all $\theta \in \Omega_0$. Hence, there is a need to strike an appropriate balance between the two goals of low power in 0 and high power in 1.

The most popular method for striking a balance between the two goals is to choose a number $\alpha 0$ between 0 and 1 and require that

$\pi(\theta|\delta) \leq \alpha 0$, for all $\theta \in \Omega_0$. (9.1.6)

Then, among all tests that satisfy (9.1.6), the statistician seeks a test whose power function is as high as can be obtained for $\theta \in 1$. This method is discussed in Sections 9.2 and 9.3. Another method of balancing the probabilities of type I and type

II errors is to minimize a linear combination of the different probabilities of error.

TESTING SIMPLE HYPOTHESIS

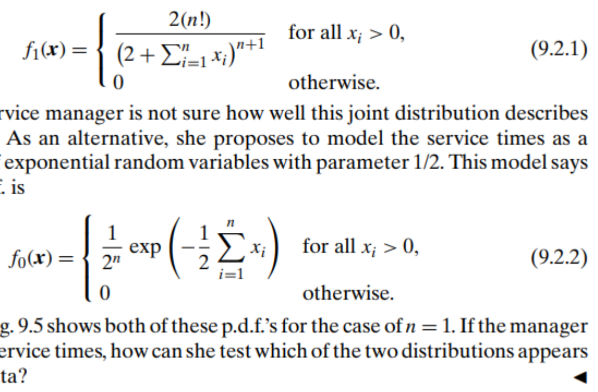
The simplest hypothesis-testing situation is that in which there are only two possible

values of the parameter. In such cases, it is possible to identify a collection of test procedures that have certain optimal properties.

Introduction

Service Times in a Queue. In Example 3.7.5, we modeled the service times X =

(X1,...,Xn) of n customers in a queue as having the joint distribution with joint p.d.f.



$$f_1(\boldsymbol{x}) = \begin{cases} \dfrac{2(n!)}{\left(2+\sum_{i=1}^{n} x_i\right)^{n+1}} & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \qquad (9.2.1)$$

Suppose that a service manager is not sure how well this joint distribution describes the service times. As an alternative, she proposes to model the service times as a random sample of exponential random variables with parameter 1/2. This model says that the joint p.d.f. is

$$f_0(\boldsymbol{x}) = \begin{cases} \dfrac{1}{2^n} \exp\left(-\dfrac{1}{2}\sum_{i=1}^{n} x_i\right) & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \qquad (9.2.2)$$

For illustration, Fig. 9.5 shows both of these p.d.f.'s for the case of $n = 1$. If the manager observes several service times, how can she test which of the two distributions appears to describe the data? ◀

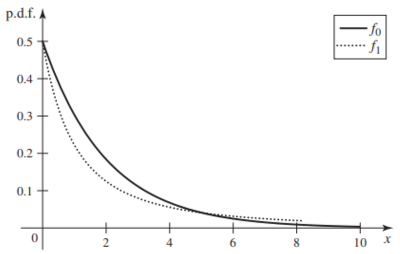


Figure 9.5 Graphs of the two competing p.d.f.'s in Example 9.2.1 with $n = 1$.

H0: $\theta = \theta_0$,

H1: $\theta = \theta_1$. In this case, 0 = $\{\theta_0\}$ and 1 = $\{\theta_1\}$ are both singleton sets.

For the special case in which X is a random sample from a distribution with

univariate p.d.f. or p.f. f (x|θ), we then have, for i = 0 or i = 1,

fi(x) = f (x1|θi)f (x2|θi) ... f (xn|θi).

The Two Types of Errors

When a test of the hypotheses (9.2.3) is being carried out, we have special notation for the probabilities of type I and type II errors. For each test procedure δ, we shall let $\alpha(\delta)$ denote the probability of an error of type I and shall let $\beta(\delta)$ denote the probability of an error of type II. Thus,

$\alpha(\delta)$ = Pr(Rejecting $H_0|\theta = \theta_0$),
$\beta(\delta)$ = Pr(Not Rejecting $H_0|\theta = \theta_1$).

Power of the Test

For each parameter vector θ = (μ1, μ2, σ2), the power function of the two-sample t test can be computed using the noncentral t distribution introduced in Definition Almost identical reasoning to that which led to Theorem Proves the following.

Power of Two-Sample t Test. Assume the conditions stated earlier in this section. Let U be defined in . Then U has the noncentral t distribution with m + n - 2 degrees of freedom and noncentrality parameter

$$\psi = \frac{\mu_1 - \mu_2}{\sigma\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}}.$$

Level of significance

Failure Times of Ball Bearings. In Example 5.6.9, we observed the failure times of 23 ball bearings, and we modeled the logarithms of these failure times as normal random variables. Suppose that we are not so confident that the normal distribution is a good model for the logarithms of the failure times. Is there a way to test the null hypothesis that a normal distribution is a good model against the alternative that no normal distribution is a good model? Is there a way to estimate features of the distribution of failure times (such as the median, variance, etc.) if we are unwilling to model the data as normal random variables?

In each of the problems of estimation and testing hypotheses that we considered statistician come from distributions for which the exact form is known, even though the values of some parameters are unknown. For example, it might be assumed that the observations form a random sample from a Poisson distribution for which the mean is unknown, or it might be assumed that the observations come from two normal distributions for which the means and variances are unknown. In other words, we have assumed that the observations come from a certain parametric family of distributions, and a statistical inference must be made about the values of the parameters defining that family.

In many of the problems to be discussed in this chapter, we shall not assume that the available observations come from a particular parametric family of distributions.

Rather, we shall study inferences that can be made about the distribution from which the observations come, without making special assumptions about the form of that distribution.

As one example, we might simply assume that the observations form a random sample from a continuous distribution, without specifying the form of this distribution any further, and we might then investigate the possibility that this distribution is a normal distribution. As a second example, we might be interested in making an inference about the value of the median of the distribution from which the sample was drawn, and we might assume only that this distribution is continuous.

As a third example, we might be interested in investigating the possibility that two

independent random samples actually come from the same distribution, and we might assume only that both distributions from which the samples are taken are continuous.

Problems in which the possible distributions of the observations are not restricted to a specific parametric family are called nonparametric problems, and the statistical methods that are applicable in such problems are called nonparametric methods.

Categorical Data

Blood Types. In Example 5.9.3, we learned about a study of blood types among a sample of 6004 white Californians. Suppose that the actual counts of people with the four blood types are given in Table 10.1. We might be interested in whether or not

These data are consistent with a theory that predicts a particular set of probabilities for the blood types. Table 10.2 gives theoretical probabilities for the four blood types.

How can we go about testing the null hypothesis that the theoretical probabilities in Table 10.2 are the probabilities with which the data in Table 10.1 were sampled?

In this section and the next four sections, we shall consider statistical problems based on data such that each observation can be classified as belonging to one of a finite number of possible categories or types. Observations of this type are called

categorical data. Since there are only a finite number of possible categories in these problems, and since we are interested in making inferences about the probabilities of

These categories, these problems actually involve just a finite number of parameters. However, as we shall see, methods based on categorical data can be usefully applied in both parametric and nonparametric problems.

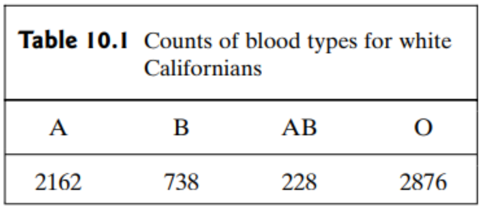


Table 10.1 Counts of blood types for white Californians

A	B	AB	O
2162	738	228	2876

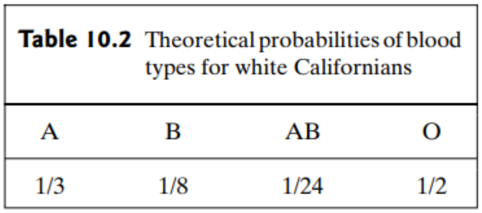


Table 10.2 Theoretical probabilities of blood types for white Californians

A	B	AB	O
1/3	1/8	1/24	1/2

The χ2 Test

Suppose that a large population consists of items of k different types, and let pi denote the probability that an item selected at random will be of type i (i = 1, . . . , k).

Example 10.1.2 is of this type with k = 4. Of course, pi ≥ 0 for i = 1,...,k and k

i=1 pi = 1. Let p0

1,...,p0 k be specific numbers such that p0

i > 0 for i = 1,...,k

and k

i=1 p^0

i = 1, and suppose that the following hypotheses are to be tested:

H_0: pi = p^0

i for i = 1, . . . , k,

H_1: pi = p^0

i for at least one value of i.

We shall assume that a random sample of size n is to be taken from the given population. That is, n independent observations are to be taken, and there is probability pi that each observation will be of type i (i = 1, . . . , k). On the basis of these

In observations, the hypotheses (10.1.1) are to be tested.

Fori = 1,...,k, we shall let Ni denote the number of observations in the random

samples that are of type i. Thus, N1,...,Nk are nonnegative integers such that k

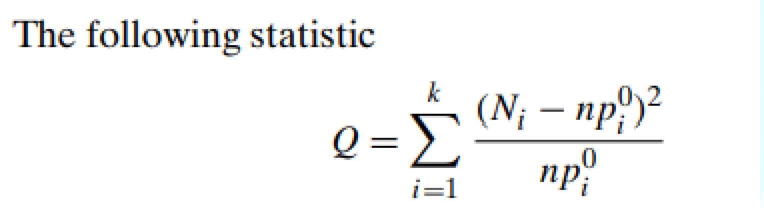
i=1 Ni = n. Indeed, (N1,...,Nn) has the multinomial distribution (see Sec. 5.9) with parameters n and p = (p1,...,pk). When the null hypothesis H0 is true, the expected number of observations of type i is np^0

i (i = 1, . . . , k). The difference

between the actual number of observations Ni and the expected number np^0

i will tend to be smaller when H_0 is true than when H_0 is not true. It seems reasonable, therefore,

to base a test of the hypotheses (10.1.1) on values of the differences Ni - np^0i for i = 1,...,k and reject H_0 when the magnitudes of these differences are relatively Large



The following statistic

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0}$$

The χ2 Test for Composite Null Hypotheses

In order to carry out a χ2 test of goodness-of-fit of the hypotheses the statistic Q defined by Eq. (10.1.2) must be modified because the expected number $np^0{}_i$of

observations of type i in a random sample of n observations is no longer completely

specified by the null hypothesis H_o. The modification that is used is simply to replace np^o

i by the M.L.E. of this expected number under the assumption that H0 is true. In other words, if θˆ denotes the M.L.E. of the parameter vector θ based on the observed
numbers N1,...,Nk, then the statistic Q is defined as follows:

$$Q = \sum_{i=1}^{k} \frac{[N_i - n\pi_i(\hat{\theta})]^2}{n\pi_i(\hat{\theta})}.$$

The χ2 Test of Independence

The χ2 tests described in Sec. 10.2 can be applied to the problem of testing the hypotheses (10.3.3). Each individual in the population from which the sample is taken must belong in one of the RC cells of the contingency table. Under the null hypothesis
H0, the unknown probabilities pij of these cells have been expressed as functions
of the unknown parameters pi+ and p+j .
Since $\sum_{i=1}^{R} p_{i+} = 1$ and $\sum_{j=1}^{C} p_{+j} = 1,$ the actual number of unknown parameters to be estimated when H0 is true is s =
(R - 1) + (C - 1), or s = R + C - 2.
For i = 1,...,R, and j = 1,...,C, let Eˆ ij denote the M.L.E., when H0 is true,
of the expected number of observations that will be classified in the ith row and thej th column of the table. In this problem, the statistic Q defined by will have the following form:

$$Q = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(N_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}.$$

Next, we shall consider the form of the estimator Eˆij .

$$\hat{E}_{ij} = n\left(\frac{N_{i+}}{n}\right)\left(\frac{N_{+j}}{n}\right) = \frac{N_{i+}N_{+j}}{n}.$$

Analysis of Variance techniques

The One-Way Layout

Example

Calories in Hot Dogs. Moore and McCabe (1999) describe data gathered by Consumer Reports. The data comprise (among other things) calorie contents from 63 brands of hot dogs. (See Table 11.15.) The hot dogs come in four varieties: beef, "meat" (don't ask), poultry, and "specialty." (Specialty hot dogs include stuffing such as cheese or chili.) It is interesting to know whether, and to to what extent, the different varieties differ in their calorie contents. Data structures of the sort in this example, consisting of several groups of similar random variables, are the subject of this section.

In this section and in the remainder of this chapter, we shall study a topic known as the analysis of variance, abbreviated ANOVA. Problems **of ANOVA** are actually problems of multiple regression in which the design matrix **Z** has a very special form. In other words, the study of ANOVA can be placed within the framework of the general linear model (Definition 11.5.1), if we continue to make the basic assumptions for such a model: The observations that are obtained are

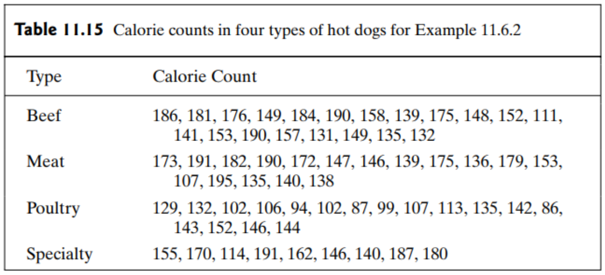


Table 11.15 Calorie counts in four types of hot dogs for Example 11.6.2

Type	Calorie Count
Beef	186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141, 153, 190, 157, 131, 149, 135, 132
Meat	173, 191, 182, 190, 172, 147, 146, 139, 175, 136, 179, 153, 107, 195, 135, 140, 138
Poultry	129, 132, 102, 106, 94, 102, 87, 99, 107, 113, 135, 142, 86, 143, 152, 146, 144
Specialty	155, 170, 114, 191, 162, 146, 140, 187, 180

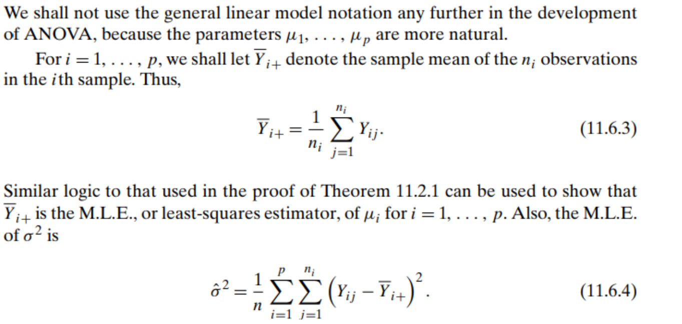
independent and normally distributed; all these observations have the same variance σ2; and the mean of each observation can be represented as a linear combination of certain unknown parameters. The theory and methodology of ANOVA were mainly

developed by R. A. Fisher during the 1920s.

We shall begin our study of ANOVA by considering a problem known as the one-way layout. In this problem, it is assumed that random samples from p different normal distributions are available, each of these distributions has the same variance

σ2, and the means of the p distributions are to be compared on the basis of the observed values in the samples. This problem was considered for two populations

(p = 2) in Sec. 9.6, and the results to be presented here for an arbitrary value of p will generalize those presented in Sec. 9.6. Specifically, we shall now make the following assumption: For i = 1,...,p, the random variables Yi1,...,Yini , form a random sample of ni observations from the normal distribution with mean μi and variance σ2, and the values of μ1,...,μp and σ2 are unknown



We shall not use the general linear model notation any further in the development of ANOVA, because the parameters $\mu_1, \ldots, \mu_p$ are more natural.

For $i = 1, \ldots, p$, we shall let $\overline{Y}_{i+}$ denote the sample mean of the n_i observations in the ith sample. Thus,

$$\overline{Y}_{i+} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}. \qquad (11.6.3)$$

Similar logic to that used in the proof of Theorem 11.2.1 can be used to show that $\overline{Y}_{i+}$ is the M.L.E., or least-squares estimator, of μ_i for $i = 1, \ldots, p$. Also, the M.L.E. of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i+}\right)^2. \qquad (11.6.4)$$

The Two-Way Layout

Radioactive Isotope in Milk. Suppose that in an experiment to measure the concentration of a certain radioactive isotope in milk, specimens of milk are obtained from four

different dairies, and the concentration of the isotope in each specimen is measured

by three different methods. If we let Y denote the measurement that is made for the
specimen from the ith dairy by using the j th method, for i = 1, 2, 3, 4 and j = 1, 2, 3, then in this example there will be a total of 12 measurements. There are two main questions of interest in this example. The first is whether the concentration of the isotope is the same in the milk of all four dairies. The second question is whether the
three different methods produce concentration measurements that appear to differ.

A problem of the type in Example 11.7.1, in which the value of the random
a variable being observed is affected by two factors, called a two-way layout. In the general two-way layout, there are two factors, which we shall call A and B. We shall assume that there are I possible different values, or different levels, of factor A, and that there are J possible different values, or different levels, of factor B.
For i = 1,...,I and j = 1,...,J , an observation Yij of the variable being studied
is obtained when factor A has the value i and factor B has the value j . If the IJ observations are arranged in a matrix as in Table 11.20, then Yij is the observation in
the (i, j) cell of the matrix.
We shall continue to make the assumptions of the general linear model for the two-way layout. Thus, we shall assume that all the observations Yij are independent, each observation has a normal distribution, and all the observations have the same
variance σ2. In this section, we specialize the assumption about the mean E(Yij) as
follows: We shall assume not only that E(Yij) depends on the values i and j of the

two factors, but also that there exist numbers θ1,...,θI and ψ1,...,ψJ such that

$$E(Y_{ij}) = \theta_i + \psi_j \quad \text{for } i = 1, \dots, I \quad \text{and } j = 1, \dots, J.$$

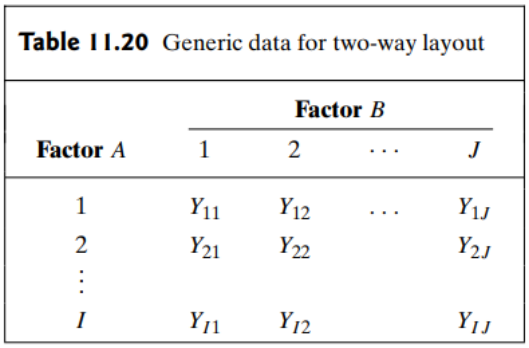
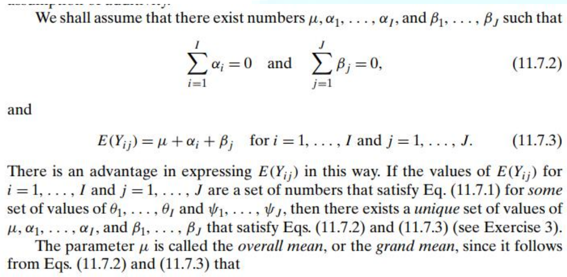


Table 11.20 Generic data for two-way layout

	Factor B			
Factor A	1	2	$\cdots$	J
1	Y_{11}	Y_{12}	$\cdots$	Y_{1J}
2	Y_{21}	Y_{22}		Y_{2J}
$\vdots$				
I	Y_{I1}	Y_{I2}		Y_{IJ}



We shall assume that there exist numbers $\mu, \alpha_1, \dots, \alpha_I$, and $\beta_1, \dots, \beta_J$ such that

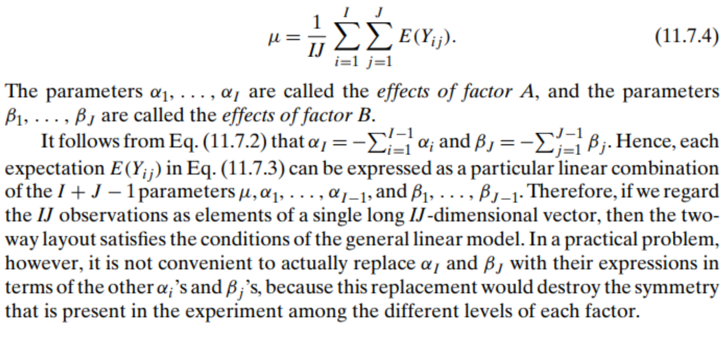
$$\sum_{i=1}^{I} \alpha_i = 0 \quad \text{and} \quad \sum_{j=1}^{J} \beta_j = 0, \tag{11.7.2}$$

and

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j \quad \text{for } i = 1, \dots, I \text{ and } j = 1, \dots, J. \tag{11.7.3}$$

There is an advantage in expressing $E(Y_{ij})$ in this way. If the values of $E(Y_{ij})$ for $i = 1, \dots, I$ and $j = 1, \dots, J$ are a set of numbers that satisfy Eq. (11.7.1) for *some* set of values of $\theta_1, \dots, \theta_I$ and $\psi_1, \dots, \psi_J$, then there exists a *unique* set of values of $\mu, \alpha_1, \dots, \alpha_I$, and $\beta_1, \dots, \beta_J$ that satisfy Eqs. (11.7.2) and (11.7.3) (see Exercise 3).

The parameter μ is called the *overall mean*, or the *grand mean*, since it follows from Eqs. (11.7.2) and (11.7.3) that



$$\mu = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} E(Y_{ij}). \tag{11.7.4}$$

The parameters $\alpha_1, \dots, \alpha_I$ are called the *effects of factor A*, and the parameters $\beta_1, \dots, \beta_J$ are called the *effects of factor B*.

It follows from Eq. (11.7.2) that $\alpha_I = -\sum_{i=1}^{I-1} \alpha_i$ and $\beta_J = -\sum_{j=1}^{J-1} \beta_j$. Hence, each expectation $E(Y_{ij})$ in Eq. (11.7.3) can be expressed as a particular linear combination of the $I + J - 1$ parameters $\mu, \alpha_1, \dots, \alpha_{I-1}$, and $\beta_1, \dots, \beta_{J-1}$. Therefore, if we regard the IJ observations as elements of a single long IJ-dimensional vector, then the two-way layout satisfies the conditions of the general linear model. In a practical problem, however, it is not convenient to actually replace α_I and β_J with their expressions in terms of the other α_i's and β_j's, because this replacement would destroy the symmetry that is present in the experiment among the different levels of each factor.