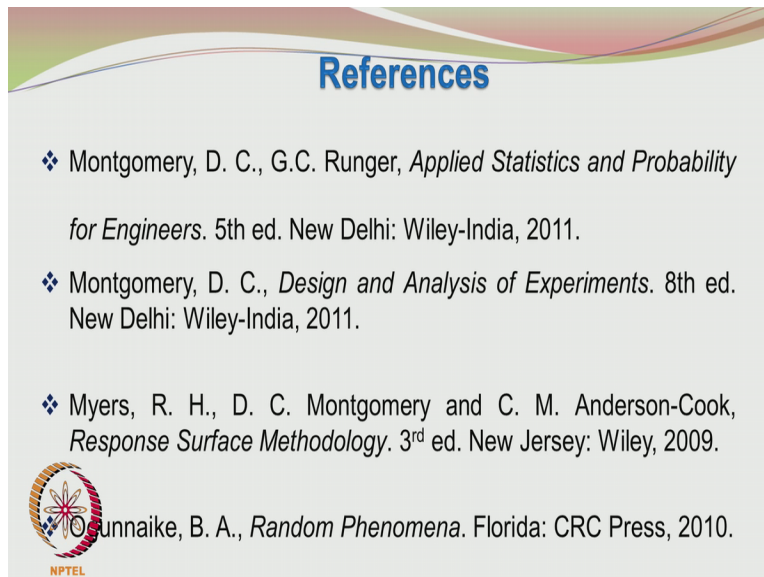


Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology - Madras

Lecture - 53
Statistics for Experimentalists - Summary Part A

In today's lecture, we will be summarizing the key issues in Statistics for Experimentalists course.

(Refer Slide Time: 00:33)

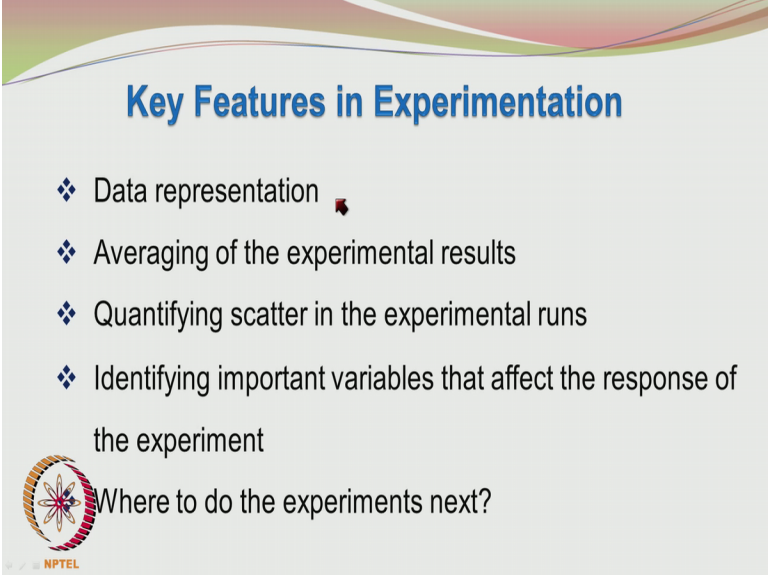


The books which are referenced here are very useful and quite reader-friendly, they also have the advantage of listing a large number of problems which may be easily worked out either with the help of a calculator or with other statistical design software. The first book would be a good reference text book prescribed text book rather for semester course on Statistics for Experimentalists, more details on Design and Analysis of Experiments are given by the second book.

The third book is slightly more advance one, but again very user-friendly, and it has lot of advanced techniques stresses more on second order designs various optimal designs, and also focuses on the Response Surface Methodology. So this is a good bridge from the fundamental learning of the course and advanced studies for those who are interested. The last book Random Phenomena also looks at things in a nice manner.

And give some interesting explanations on very difficult topics, it in fact put things in perspective and helps you to understand some difficult concepts.


(Refer Slide Time: 02:19)



Key Features in Experimentation

- ❖ Data representation
- ❖ Averaging of the experimental results
- ❖ Quantifying scatter in the experimental runs
- ❖ Identifying important variables that affect the response of the experiment

Where to do the experiments next?

 NPTEL

So the key features in experimental work are data representation, we had looked at box plots, scatter plots, and normal probability plots, histogram. So each have certain applications, and it is important that we use these diagrams or figures as frequently as possible to support our experimental findings. Whenever we do experiments we do repeats, and the important thing to notice just the average of the experimental results is not sufficient.

You also have to quantify the scatter in the experimental data, and see how much is the scatter and will it be smaller than the variation caused by your factors, or it will be comparable to the variation caused by the factors in your experiment. So that you can then make the correct conclusion. And you can do experiments in a planned manner, and unambiguously identify the important variables that affect the response of the experiment.

And after performing the experiment, the next question is where to do the experiments further, and how to go in the correct path, so that eventually the optimum conditions may be identified.

(Refer Slide Time: 04:03)

Importance of Variability

- ❖ Scatter in experimental data is but the law of nature
- ❖ It is possible to handle scattered experimental data and still draw meaningful conclusions
- ❖ In addition to **averaging** experimental data, importance must be attached to the **variability** in the data

It is important to stress upon the impact on the variability, the scatter in experimental data is unavoidable, it is nothing but the law of nature. It is possible to handle scattered experimental data and still draw meaningful conclusions with the statistical tools, and as I said earlier in addition to averaging of experimental data importance must be attached to the variability in the data.

(Refer Slide Time: 04:34)

Variability in Experimental Outcome

$$y_i = \eta_i + \varepsilon_i$$

- ❖ Linear combination of error and true but unknown process response is assumed.
- ❖ The random component (ε_i) is responsible for the observed variations when repeating the same experiment.

So the variability in the experimental outcome is modelled as a linear combination of the mean value of the response, and the random error component. So the random error component epsilon i is responsible for the observed variations when repeating the same experiment.

(Refer Slide Time: 05:00)

Random Variable

- ❖ Denoted by X and once its value is known after the experiment, it is termed as x .
- ❖ Is the genesis for probability distribution functions
- ❖ Probability distributions are used to predict the behavior of a group of randomly behaving entities



So with this background we embarked upon an exciting journey concerning with random variables, it is something where the probabilities were associated. We talked about the discrete probability mass functions. And then we also talked about for continuous random variables probability distribution functions, so we denoted the random variable as capital X , and once its value was known after an experiment or a sample survey it is represented as small x .

The random variable X is the originator for probability distribution functions, and probability distributions are used to predict the behaviour of a group of randomly behaving entities. You cannot predict a single randomly behaving person's behavior, but when you have a collection of such people their actions maybe better quantified.

(Refer Slide Time: 06:19)

Probability Density Function: interpretation

The probability density function describes the distribution of probabilities in the continuous random variable domain.



So we talked about probability distributions, and the probability distribution function describes the distribution of probabilities in the continuous random variable domain. For discrete random variables we talked about probability mass function, and you assigned the weight for a probability for every discrete outcome in the discrete random variable domain in the sample space, which contains the set of possible outcomes from a particular experiment.

And you assigned the probability to each of them, and when you add up all the probabilities they stood total 1, but discrete random variables are not that frequently encountered, we more commonly encounter continuous random variables, and we describe the behavior of the random variables through appropriate probability density functions. And we calculate the probability of a random variable taking value between A and B for instance using the appropriate distribution.

(Refer Slide Time: 07:31)

Definition

The probability distribution function $f(x)$ for continuous random variables is defined as

$$P(X \leq z) = \int_{-\infty}^z f(x)dx = F(z)$$

$F(z)$ is the cumulative distribution function for a continuous random variable X .



So we defined the probability distribution function f of x for continuous random variables as probability of X taking a value $\leq z$ then we go from $-\infty$ to $+z$ the value mentioned here, and then this is the probability density function form it is integrated, and we get the cumulative distribution functions for a random variable X , and that is represented as F of z .

(Refer Slide Time: 07:59)

Features of Probability Density Functions

What is the probability of x lying between values ' a ' and ' b '?

$$\begin{aligned} P(a \leq X \leq b) &= \int_a^b f(x)dx \\ &= \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx = F(b) - F(a) \end{aligned}$$



So if you want to find the probability of random variable lying between a and b , it is the presented as a to b integral f of x dx , and that is the cumulative distribution at b -cumulative distribution at a .

(Refer Slide Time: 08:19)

Mean and Variance of Probability Density Function

The mean for the continuous distribution is given by

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

The variance of the discrete probability distribution is

$$\sigma^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$$



Then we describe some important features of the continuous probability distribution function, we talked about the mean and variance of the distribution, and the μ = expected value of x = $-\infty$ to $+\infty$ $x f(x) dx$. The variance of the probability distribution function was given as $E[(X - \mu)^2]$ = $-\infty$ to $+\infty$ $(x - \mu)^2 f(x) dx$, there is a typo here I will correct it. So the variance of the continuous probability distribution is given by $\sigma^2 = E[(X - \mu)^2]$ = $-\infty$ to $+\infty$ $(x - \mu)^2 f(x) dx$.

(Refer Slide Time: 09:07)

Normal Probability Density Functions

- ❖ Finds several applications in science and engineering
- ❖ Many distributions such as the student t distribution tend to the normal distribution with increasing degrees of freedom
- ❖ The parameters of this distribution are mean (μ) and standard deviation (σ) directly.

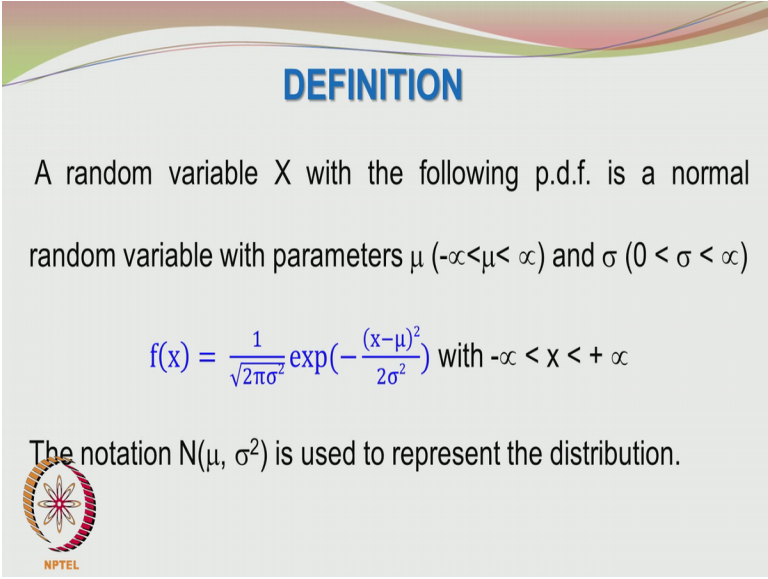


Then we talked in length about the normal probability density functions, this is very commonly encountered in our statistical applications, we are going to discuss about the central limit theorem. And then when you look at various distributions like the T-distribution or the Chi-

square distribution, they would tend to normal distributions under certain special conditions. And so we frequently resort to this normal probability density function or the Gaussian distribution.

And one advantage of this normal probability distribution is the mean and variance μ and σ^2 are themselves the parameters of the distribution.

(Refer Slide Time: 10:00)




DEFINITION

A random variable X with the following p.d.f. is a normal random variable with parameters μ ($-\infty < \mu < \infty$) and σ ($0 < \sigma < \infty$)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ with } -\infty < x < +\infty$$

The notation $N(\mu, \sigma^2)$ is used to represent the distribution.



So the probability density function for the normal distribution is given by $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ with $-\infty < x < +\infty$. We also talked about the lognormal distribution, so as far as normal distribution is concerned, we use a notation $N(\mu, \sigma^2)$ to represent the distribution.

(Refer Slide Time: 10:34)

Standardizing a Normal Random Variable

A normal random variable (X) with mean (μ) and standard deviation (σ) may be converted into a standard normal random variable by the transformation



$$Z = \frac{X - \mu}{\sigma}$$

There can be different normal distributions for different means and variances different μ and different σ^2 , it will be good if we can standardize them or normalized them such that 1 probability distribution graph or table is sufficient to calculate the required probabilities. Otherwise, for each μ and σ^2 , we may require different probability tables.

So one simple way to do it is to convert the random variable X which is following the normal distribution into a standard normal variable such that the standard normal variable refers to normal distribution of mean 0 and variance 1. So how to do it? We take the usual normal variable X and then we subtract the mean μ from it and divide by σ . Once we do that we call the transformed variable as $z = (x - \mu) / \sigma$.

This follows a standard normal distribution that means it has a mean of 0 and variance of 1.

(Refer Slide Time: 11:57)

Standardizing a Normal Random Variable

Hence this transformed random variable X has a zero mean and variance 1 and hence is a standard normal random variable.

The cumulative distribution of a standard normal random variable is denoted as $\Phi(z) = P(Z \leq z)$

So the cumulative distribution of a standard normal random variable is denoted as Φ of z =probability of capital $Z \leq z$.

(Refer Slide Time: 12:05)

Population

We need to understand the characteristic features of populations from a decision making, quality control or marketing point of view.

Understanding the population helps us to set our goals, objectives, process settings etc.

Then we talked at length about populations, the populations have to be understood from a decision making quality control or marketing points of view, and once we understand the population we can set our goals objectives process settings etc.

(Refer Slide Time: 12:27)

SAMPLE

Since questioning or testing each entity in the population is not practical, we need to **sample** the population and get to know from the sample attributes

estimates of the population mean, variance, nature of distribution etc.



So we cannot collect the data from the entire population, so we have to make use of sampling we talked about random samples which hopefully will give us sufficient information about the population parameters. So we need to get to know from the sample attributes estimates of the population mean variance nature of the distribution etc.

(Refer Slide Time: 12:56)

Population and Random Samples

The sample should comprise of independent observations which are coming from the same population.

In other words, they should represent the same probability distribution.




The sample should comprise of independent observations which are coming from the same population. In other words, they should represent the same probability distribution.

(Refer Slide Time: 13:09)

Population and Random Samples

- ❖ Sampled elements must be independent of each other
- ❖ Each element in the sample should have equal chances of being picked
- ❖ More number of sample elements, more confident and precise we feel about the responses




The sampled elements must be independent of each other, and each sample element should have an equal probability of getting picked, and the size of the sample is very important higher the sample size the more we feel confident about the precision of the parameters.

(Refer Slide Time: 13:32)

Population and Random Samples

The properties of random samples (i.e. sample mean, sample variance etc.) may be treated as random variables themselves. These are functions of random variables and termed as **statistics**.



Once you have random sample, we can use the collected data to find the sample mean and sample variance, and the sample mean and sample variance are themselves functions of the random variables, and hence they are random variables themselves. And these functions are also called as statistics, hence the sample mean \bar{X} and sample variance S^2 are called as statistics.

(Refer Slide Time: 14:01)

Sample Mean (\bar{X}) and Variance (S^2)

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

❖ Here n is the **sample size**.

❖ In definition of S^2 , not all X_i are independent as the



constraint $\sum_{i=1}^n (X_i - \bar{X}) = 0$ has to be satisfied.

Now how do we define the sample mean we found that \bar{X} is given by $\sum_{i=1}^n X_i / n$. $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$, here n is the size of the sample. As far as the sample variance goes we are dividing it by $n-1$, because not all the deviations from the mean are independent of each other, we have only $n-1$ deviations from the mean which are independent. And hence $\sum_{i=1}^n (X_i - \bar{X}) = 0$, so this $n-1$ may also be termed as degrees of freedom in some of the variance calculations as we saw in the lectures.

(Refer Slide Time: 14:51)

Point Estimators

The statistics

❖ sample mean \bar{X}

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

❖ sample variance (S^2)

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

are the point estimators of the unknown parameters μ and σ^2



respectively.

So we looked at point estimators, point estimators mean that we are going to get single values of the population parameters, and we looked at sample mean \bar{X} and sample variance S^2 . One important advantage of this point estimators is that they are unbiased estimators, expected

value of \bar{X} will be $=\mu$, and it can be shown that expected value of $S^2 = \sigma^2$ the population variance itself.

So from the sample mean \bar{X} , we are going to estimate point estimate about this population mean μ and from the sample variance S^2 , we are getting an idea or a point estimate about σ^2 the population variance.

(Refer Slide Time: 15:45)

Point Estimators

\bar{X} and S^2

are **point estimates**

of population parameters

μ and σ^2 respectively

NPTEL

So we call \bar{X} and S^2 are point estimates of the population parameters μ and σ^2 respectively, let me make that correction. So we are now looking at point estimators \bar{X} and S^2 are point estimates of the population parameters μ and σ^2 respectively.

(Refer Slide Time: 16:06)

Sampling Distributions

- ❖ The statistics \bar{X} and S^2 also have a probability distribution associated with each of them
- ❖ They are referred to as the **sampling distributions** of the sample mean and sample variance respectively



Since statistics are also random variables \bar{X} and S^2 also have a probability distribution associated with each of them, and they are referred to as the sampling distribution of the sample mean and sample variance respectively. Any random variable has a probability distribution associated with it, and so do \bar{X} and S^2 . What is really the meaning of the probability distribution associated with the 2 sample statistics?

So if you take multiple random samples, each sample mean gives a different sample mean, each sample make give a different sample variance. So the distribution of the sample means and the sample variances will constitute the sampling distribution of the means and sampling distribution of the variances. So each of these statistics have a probability distribution associated with it.

(Refer Slide Time: 17:22)

Properties of the Sampling Distribution

We will look at a general case involving n **independent** random variables. However, we shall assume that all of them have come from populations that have the same mean μ and variance σ^2 .



So what are the properties of the sampling distribution, we will look at the general case involving n independent random variables. However, we will assume that all of them come from populations that have the same mean μ and variance σ^2 . So the condition of independence of the random variables is important.

(Refer Slide Time: 17:44)

Sampling Distribution of the Mean

Since $E(X) = \mu$, and $E(X_1) = E(X_2) = \dots = E(X_n) = \mu$
(n times μ for n random variables). This simply becomes

$$E(\bar{X}) = \frac{E(X_1) + E(X_2) + \dots + E(X_n)}{n}$$

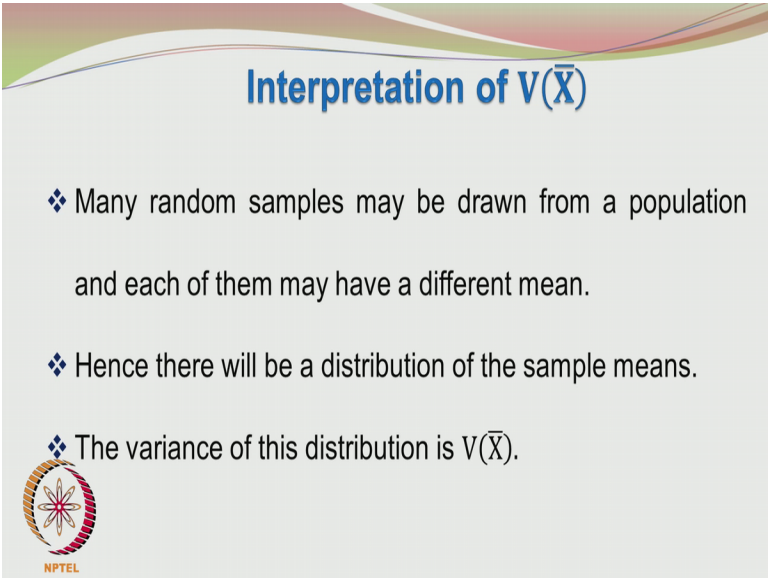
$$E(\bar{X}) = \frac{n\mu}{n} = \mu$$



So when we look at the expected value of \bar{X} , it would be expected value of X_1 + expected value of X_2 + so on to expected value of X_n/n , so expected value of \bar{X} would be $n\mu/n$ because each of the random variables are coming from the populations with the same mean μ and variance σ^2 . So expected value of X_1 would be μ expected value of X_2 will be


μ and so on to expected value of X_n will be μ , so you will have expected value of $\bar{X} = \mu$ itself.

(Refer Slide Time: 18:20)



Interpretation of $V(\bar{X})$

- ❖ Many random samples may be drawn from a population and each of them may have a different mean.
- ❖ Hence there will be a distribution of the sample means.
- ❖ The variance of this distribution is $V(\bar{X})$.



And on similar lines we can also so that expected value of $S^2 = \sigma^2$, so what is the interpretation of the variance of \bar{X} , variance of the sample mean. So many random samples can be drawn from a population and each of them have a different mean, this is understandable. So there will be a distribution of the sample means, what it means is if you plot the different sample means you have obtained in form of a frequency diagram.

You will find that certain sample mean values are more popular than the rest, so they will be occurring more frequently. This is typical of any probability distribution, because around the mean value you will get higher values of the probability distribution function. So anyway you will be having a range of sample mean values, and they will be described a probability distribution, and the spread of this sampling distribution of the means is characterized by the variance of \bar{X} .

(Refer Slide Time: 19:48)

Properties of the Sampling Distribution

If the variance of the random variable X is $V(X)$ and is equal to σ^2 ,

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = V\left(\frac{X_1}{n}\right) + V\left(\frac{X_2}{n}\right) + \dots + V\left(\frac{X_n}{n}\right)$$

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = \left(\frac{\sigma^2}{n^2}\right) + \left(\frac{\sigma^2}{n^2}\right) + \dots + \left(\frac{\sigma^2}{n^2}\right) = \left(\frac{n\sigma^2}{n^2}\right) = \frac{\sigma^2}{n}$$



So if the variance of a random variable X is variance of X and is equal to sigma squared, what is variance of \bar{X} ? It would be sigma square/n. So if sigma squared is the variance of the population, the variance of the sampling distribution \bar{X} is sigma square/n, the population probability distribution function will have a variance sigma squared, whereas the sampling distribution variance will be sigma square/n.

And how we got it? variance of \bar{X} would be variance of X_1/n + variance of X_2/n so on to variance of X_n/n , and variance of X_1/a constant would be that variance of $X_1/\text{square of the constant or sigma square/n squared}$. Since all of them are having come from populations with the same mean and same variance sigma squared, variance of X_1 will be squared, variance of X_2 will be also sigma squared so on 2 variance of X_n which will also be sigma squared.

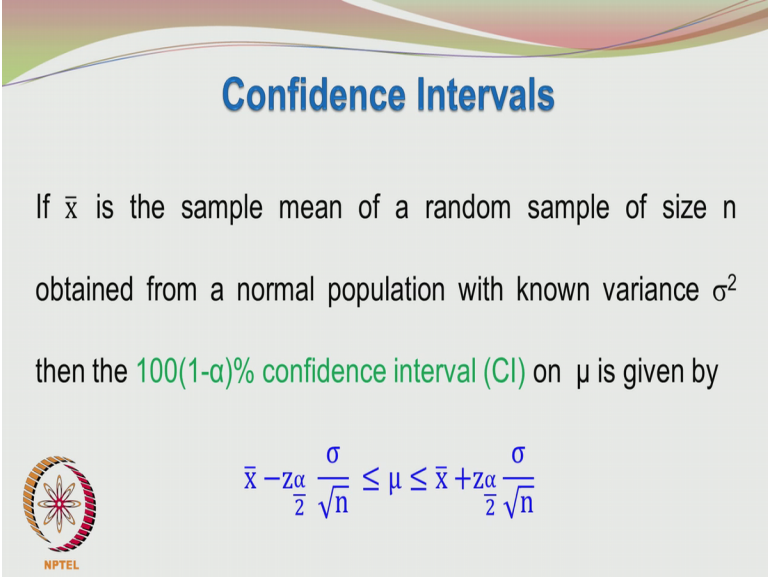
So you have n sigma squared/n squared, so variance of \bar{X} would be sigma square/n. So you are having a probability distribution of the sample means, which is nothing but a distribution of the probabilities of \bar{X} , so this distribution will have a mean same as that of population mean μ , so if the population is centered around the mean μ , the probability distribution function for the sample mean will also be centered around μ or center at μ .

So both the sample distribution as well as the population distribution have the same mean μ , the variance of the population parent population probability distribution is sigma squared. But on

the other hand, the sampling distributions variance will be σ^2/n , so when compared to the population the sampling distribution of the means will also have the same mean μ , but it will have a lower variance given by σ^2/n , where $n > 1$.

So the sampling distribution of the mean will have a variance which is $1/n$ th of the variance of the populations probability density function.


(Refer Slide Time: 22:51)



Confidence Intervals

If \bar{x} is the sample mean of a random sample of size n obtained from a normal population with known variance σ^2 then the $100(1-\alpha)\%$ confidence interval (CI) on μ is given by

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

 NPTEL

Then we looked at confidence intervals just as we had point estimates of the population parameters, we also have interval estimates, we looked at putting upper and lower bounds on the population parameters. So we defined a 95% confidence interval and that is given by $\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. Here, it is assumed that the parent population's variance σ^2 or a standard deviation σ is known as to us.

Alpha is the level of significance, since we are looking at a lower bound and upper bound on μ we use $z_{\alpha/2}$.

(Refer Slide Time: 23:46)

Confidence Intervals

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Here $z_{\alpha/2}$ is the upper 100 $\alpha/2$ percentage point of the standard normal distribution.



So what is $z_{\alpha/2}$? It is the upper 100 $\alpha/2\%$ point of the standard normal distribution.

(Refer Slide Time: 23:55)

Sampling Distributions and the Central Limit Theorem

- ❖ Even if the original probability distribution of the population is not Gaussian, the sample mean distribution tends towards the normality provided the sample size is high (say > 30).
- ❖ The sample distribution is nearly normal with mean μ and variance σ^2/n



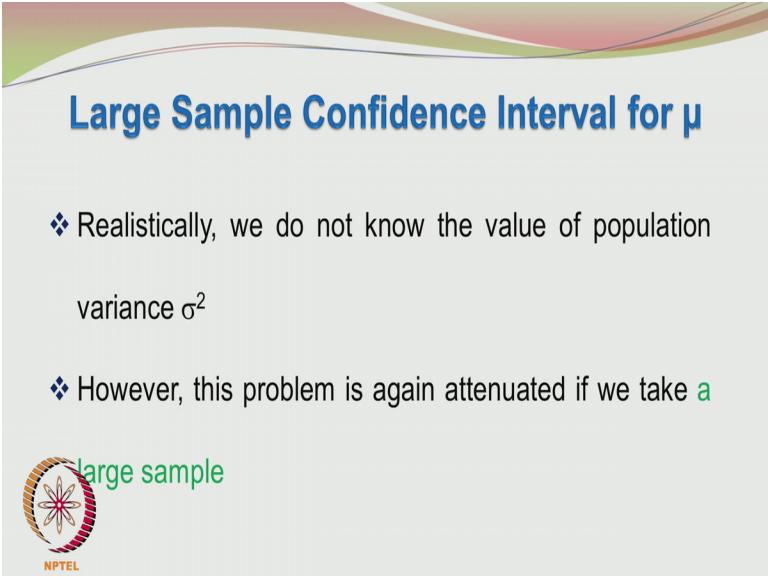
So now let us look at sampling distributions on the central limit theorem, the central limit theorem is a very useful concept or helping hand in statistics. So the population can be described by any arbitrary probability density function, even though most population tend towards normality, there may be certain populations which are described differently. So we take a sample from such a population.

If the sample size is >30 , then the sampling distribution of the mean would tend towards normality, so this is very good. Irrespective of the sample size, if the parent population is

normally distributed, then the random samples taken out of such a distribution also would be normally distributed, the random samples taken out of the population also would tend to be normally distributed irrespective of sample size.


If the parent population is not normal but we take samples of size >30 , then the sampling distribution of the sample means would be normal or tend towards normality. So this is a very useful tool in statistical analysis. So what would be the variance of such a distribution? If the sample size is >30 the sampling distribution of the mean would be distributed according to the normal distribution with the mean μ and the variance σ^2/n , μ and σ^2 are the mean and the variances of the parent population.

(Refer Slide Time: 26:08)



Large Sample Confidence Interval for μ

- ❖ Realistically, we do not know the value of population variance σ^2
- ❖ However, this problem is again attenuated if we take a large sample

 large sample

So we do not know the value of σ^2 , we also do not know the value of μ , and on top of it we do not know the σ^2 . And this problem is attenuated if we take a large sample.

(Refer Slide Time: 26:26)

Large Sample Confidence Interval for μ

We do not know σ and hence cannot define the standard normal Z . However, we have taken a large sample and hence we can replace S for σ and define Z as

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$



So when we do not know the value of sigma and we have taken a large sample, then we know that the sample is going to behave normally according to the central limit theorem, we can define a standard normal variable according to $z = (\bar{X} - \mu) / (S/\sqrt{n})$, we knock off sigma which we do not know, and in place put the sample standard deviation here, and then we divide by \sqrt{n} .

(Refer Slide Time: 26:59)

Two Sided Interval Estimates When σ^2 is Unknown

The large sample confidence interval is now defined for μ

$$P\left(\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Hence we get the $100(1-\alpha)\%$ confidence interval on μ as

$$\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}}$$



And in such cases the large sample confidence interval is now defined for μ as probability of $\bar{X} - z_{\alpha/2} S/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2} S/\sqrt{n} = 1 - \alpha$, and in from such a definition we can get a $100 \cdot (1 - \alpha)\%$ confidence interval on μ as $\bar{X} - z_{\alpha/2} S/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2} S/\sqrt{n}$.

(Refer Slide Time: 27:29)

Confidence Interval Estimates When σ^2 is Unknown

Here, for the central limit theorem to hold, the sample size should be **preferably 40 or more**, as there is additional spread due to the unknown σ and we are replacing it with



S when defining the standard normal.

When sigma squared is not known and we are using S squared instead of sigma squared, then for the central limit theorem to hold we recommend a sample size of >40 , so this is to account for the additional variability due to the unknown sigma okay.

(Refer Slide Time: 27:53)

The T- Distribution

- ❖ Used when the **sample size is small** and variance (σ^2) is unknown (as also the other parameter μ)
- ❖ The assumption made is that the population from where the sample is drawn is normal



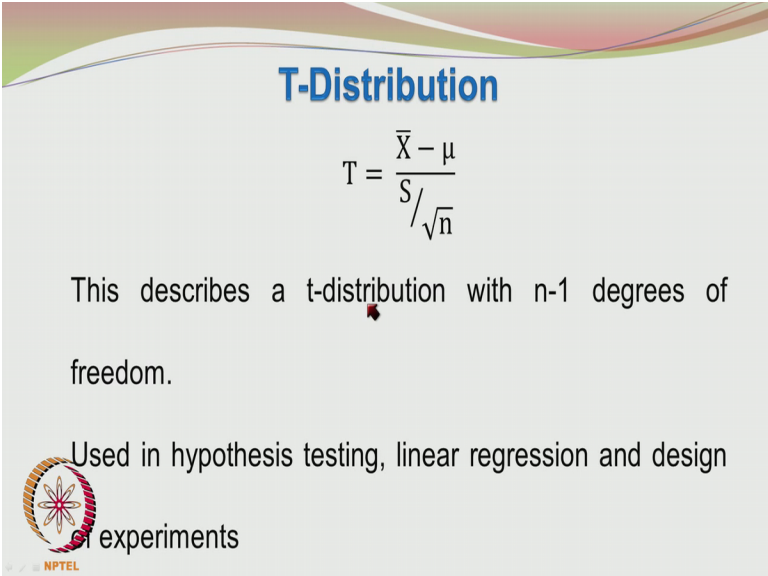
Small deviations from normality assumption is not serious.

Now we know that the population is normal and variance is known, then if the samples taken from such a normal population of known variance sigma squared are even small size sized ones, then the sampling distribution of the means will be still normal. We are taking small samples out of such a population miraculously the sigma squared off that population is known, then the sampling distribution of the means would be normal with mean μ and variance σ^2/n , there is no problem with it.

What if the parent population is normal, but the variance sigma squared is not known and the sample size is small, then of course without knowing the sigma squared we have to use S squared, but we cannot use the central limit theorem, and say that the resulting distribution would be normal. Because here sigma square is not known, and also the sample size is very small, in such cases we have to use special distribution called as T-distribution.

And the T-distribution depends upon the sample size, and n -1 is referred to as the degrees of freedom of the T-distribution, the T-distribution tends towards the normal distribution as the sample size increases towards infinity.

(Refer Slide Time: 29:52)




T-Distribution

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

This describes a t-distribution with n-1 degrees of freedom.

Used in hypothesis testing, linear regression and design of experiments



NPTEL

So we describe the T-distribution as $\bar{X} - \mu / S / \sqrt{n}$, and this T-distribution is very useful it is helpful in hypothesis testing, linear regression, analysis and design of experiments.

(Refer Slide Time: 30:12)

Chi-Square Distribution

We may require to report confidence intervals on the population variance. It may be done subject to the following assumption

Population is normally distributed



Next we look at the Chi-square distribution as such the chi-square distribution is not very frequently encountered, but the ratio of 2 chi-square distributions leads to another distribution called as F-distribution from that point of view we need to understand what is meant by the chi-square distribution. And the chi-square distribution is used whenever we need to confidence intervals on the population variances, we assume that the parent population is normally distributed.

(Refer Slide Time: 30:43)

Chi-Square Distribution

Let X_1, X_2, \dots, X_n be a random sample from a normal distribution of mean μ and variance σ^2 . S^2 is the variance of the sample. The following random variable follows a chi-square distribution with $n-1$ degrees of freedom



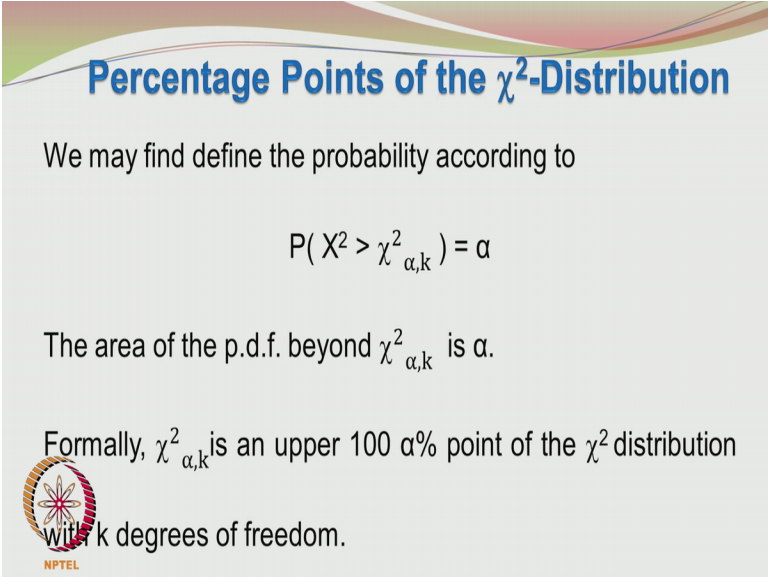
$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

And let us know define the chi-square distribution, let us say that you have sample X_1, X_2, \dots, X_n so on to X_n , which is the random sample from a normal distribution of mean μ and variance σ^2 . S^2 is the variance of the sample, the following random variable gives the or defines

the chi-square distribution with $n-1$ degrees of freedom, so let us define chi-square random variable as $n-1$ S squared/sigma squared.

This is also a function of the random variable, and hence it also has a probability distribution, the probability distribution is called as chi-square distribution.

(Refer Slide Time: 31:31)




Percentage Points of the χ^2 -Distribution

We may find define the probability according to

$$P(X^2 > \chi^2_{\alpha,k}) = \alpha$$

The area of the p.d.f. beyond $\chi^2_{\alpha,k}$ is α .

Formally, $\chi^2_{\alpha,k}$ is an upper 100 $\alpha\%$ point of the χ^2 distribution with k degrees of freedom.



And whenever we want to find probability using the chi-square distribution, we say that probability of chi-square $>$ chi-square α $k=\alpha$, where the area of the probability distribution function beyond the chi-square α k is α . So formally chi-square α k is an upper 100 $\alpha\%$ point of the chi-square distribution with k degrees of freedom.

(Refer Slide Time: 32:00)

Scope of Hypothesis Testing

Hypothesis testing concerns with the parameters of the probability distribution of the population **and not with the sample.**



However **it relies on the data from the sample** from the population of interest

So after defining the chi-square distribution it will be useful to talk about the hypothesis testing, in the hypothesis testing what do you test for? We test for queries regarding the parameters of the population okay, we are talking about μ or we are talking about σ^2 . In hypothesis testing we are using the sample attributes like sample mean and sample variance, but we are postulating or passing judgement on the population parameters μ and σ^2 , this is to be noted.

We are using the samples taken from the population, and the samples attributes obtained there from such as \bar{X} and S^2 in our analysis, but we are passing judgements on actually μ and σ^2 .

(Refer Slide Time: 33:00)

Formulation of Hypothesis

- ❖ There are two hypotheses: a. **Null** and b. **Alternate**

Hypothesis

- ❖ In defining two Hypotheses, we imply that the rejection of the

Null means automatic acceptance of its alternate.



There are 2 hypothesis here, one is null hypothesis, and other is the alternative hypothesis. And in defining the 2 hypothesis we implied that the rejection of the null hypothesis means automatic acceptance of its alternate, so the null hypothesis is usually a statement representing the status quo okay, the new process you are suggesting is not going to produce an improvement. The regression equation you are talking about none of the regression parameters are going to be influential in modelling the response.

So the status quo is to be maintained in the null hypothesis in the alternate hypothesis, we try to negate the statement made in the null hypothesis either we negated in such a way that we see it is different from a certain postulated value or we say this $>$ or $<$. So each of it will have a special kind of test, if the alternate hypothesis is saying not equal to or different than that we have to conduct what is called as the two-tailed test.

If it is altered the hypothesis, which is having a statement such as $>$ or $<$ then we have to do what is called as a one-tailed test.

(Refer Slide Time: 34:33)

Formulation of Hypothesis

We use the sample data and identify a **test statistic** (which is a function of the sample measurements) using which we try to establish the null hypothesis or its alternate and subsequently make a decision



So what we do is we use the sample data and identify a test statistic which is based on the sample measurements, and we use this to establish whether the null hypothesis or its alternate is correct. (Refer Slide Time: 34:52)

Errors in Decision Making

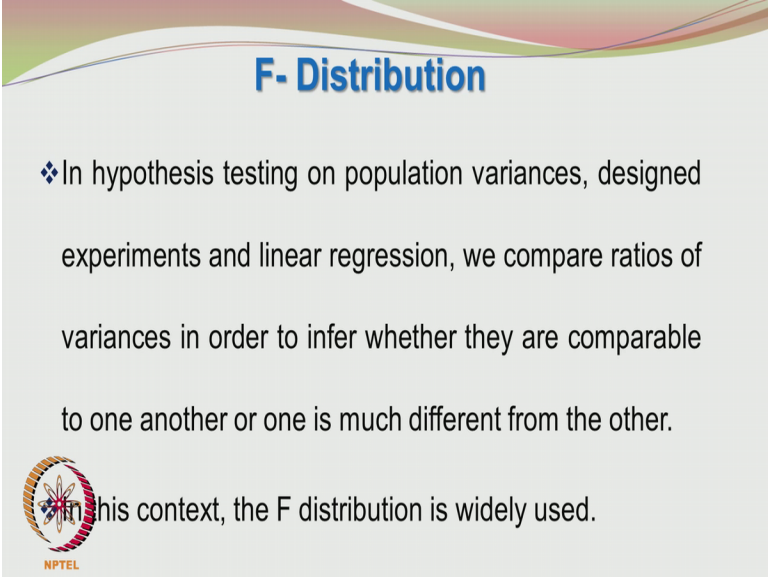
Statistical Decision	True State of Null Hypothesis	
	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision



So what are the different types of errors, we may encounter in decision making. One decision maybe accepts H_0 or do not reject H_0 , when H_0 is true it is the correct decision, when H_0 is false it is called as I type 2 error, so that means wrongly accepting the null hypothesis when it is false is called as type 2 error. And then we have what is called as a type 1 error, which is more serious that type 1 error occurs when actually H_0 is true, but you are rejecting the null hypothesis.

Then it is called as type 1 error, it is considered to be a quite a serious error. It is like saying that the null hypothesis of a Court judge being the defendant is innocent, so that is null hypothesis and eventually based on the arguments put by the prosecution he concludes that the defendant is guilty, even though he is innocent then he is committing a serious error, he is actually sentencing a innocent defendant. So wrongly rejecting the null hypothesis is called as the type 1 error. And when you reject H_0 when it H_0 is false it is a correct decision.

(Refer Slide Time: 36:18)



F- Distribution

❖ In hypothesis testing on population variances, designed experiments and linear regression, we compare ratios of variances in order to infer whether they are comparable to one another or one is much different from the other.

In this context, the F distribution is widely used.

NPTEL

Now we come to the F-distribution and the F-distribution is used in hypothesis testing linear regression and so on. And in F-distribution test we compare ratios of variances in order to infer whether they are comparable to one another or they are much different from one another. So to compare 2 variances we require the usage of the F-distribution concept.

(Refer Slide Time: 36:47)

F-Distribution

The assumptions made when comparing the variances are

- a. The two populations from which the variances were taken for comparison are both normally distributed
- b. Both the population means (μ_1, μ_2) and standard deviations (σ_1, σ_2) are unknown.



So we assume that the 2 population from which the variances were taken for comparison are both normal distributed, and both the population means μ_1, μ_2 , and standard deviations σ_1, σ_2 are not known.

(Refer Slide Time: 37:02)

Definition

The F random variable F is defined as the ratio of two independent chi-square random variables (CD_1 and CD_2) each scaled with its own degree of freedom.



$$F = \frac{CD_1/m_1}{CD_2/m_2}$$

The F random variable or the fisher random variable is defined as the ratio of the 2 independent chi-squared random variables, CD_1 and CD_2 is scaled with its own degree of freedom, so F is a ratio of $CD_1/m_1/CD_2/m_2$, CD_1 represents the first chi-square distribution with m_1 degrees of freedom, and CD_2 refers to the second chi-square distribution with m_2 degrees of freedom.

(Refer Slide Time: 37:34)

Ratio of Variances

The sample variances from these two populations are S_1^2 and S_2^2 . The two sample sizes are m and n .

The associated degrees of freedom are $m-1$ and $n-1$.



$$F = \frac{(m-1)S_1^2/[\sigma_1^2(m-1)]}{(n-1)S_2^2/[\sigma_2^2(n-1)]} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

And from the definitions for the chi-square distribution we can show that $F = S_1^2/\sigma_1^2 / S_2^2/\sigma_2^2$, and this F distribution is having $m-1$ numerator degrees of freedom and $n-1$ denominator degrees of freedom, here $m-1$ and $n-1$ are degrees of freedom for the chi-square distribution in the numerator and in the denominator.

(Refer Slide Time: 38:06)

Percentage Point

The percentage point of the F distribution is defined such that

$$P(F > f_{\alpha, m_1, m_2}) = \int_{f_{\alpha, m_1, m_2}}^{\infty} f(x) dx = \alpha$$

The percentage points in the lower tail may be found from



that of the upper tails as $f_{1-\alpha, m_1, m_2} = \frac{1}{f_{\alpha, m_2, m_1}}$

And the percentage point of the F distribution is defined such that probability of $F > f_{\alpha, m_1, m_2} = \int_{f_{\alpha, m_1, m_2}}^{\infty} f(x) dx = \alpha$. So if you have f_{α, m_1, m_2} then by taking the reciprocal of that you can find the $f_{1-\alpha, m_1, m_2}$, this is a useful result. Now we slowly move into the design of experiments after having a statistical background, the

statistical background which was presented is adequate for understanding the statistical design concepts.

It is necessary for further understanding of the experimental design concepts, so without knowing these it is not a good idea to venture straightaway into design of experiments, we need to know what is meant by the random variable, we need to know what is meant by normal probability distributions, how to compute the probabilities in such distributions, what is meant by the central limit theorem, what is meant by point estimator.

What is meant by an interval estimate, what is meant by a 95% confidence interval, so all these things are very essential. Then we also need to know about chi-squared and F-distribution, even though there are a lot of distributions here like the beta distribution, gamma distribution, Weibull distribution and so on. We do not have to learn all of them, if you learn the normal distribution T-distribution, chi-square distribution and F-distribution, it is sufficient.

Even the lognormal distribution occasionally crops up, so it is a good idea to know about it, beyond it we do not need to look into all the statistical distributions. And once you have understood a statistical distribution, you should also learn how to find the mean and variance of such distributions. Because they are very important parameters, so the samples which are drawn from such populations also have distributions okay.

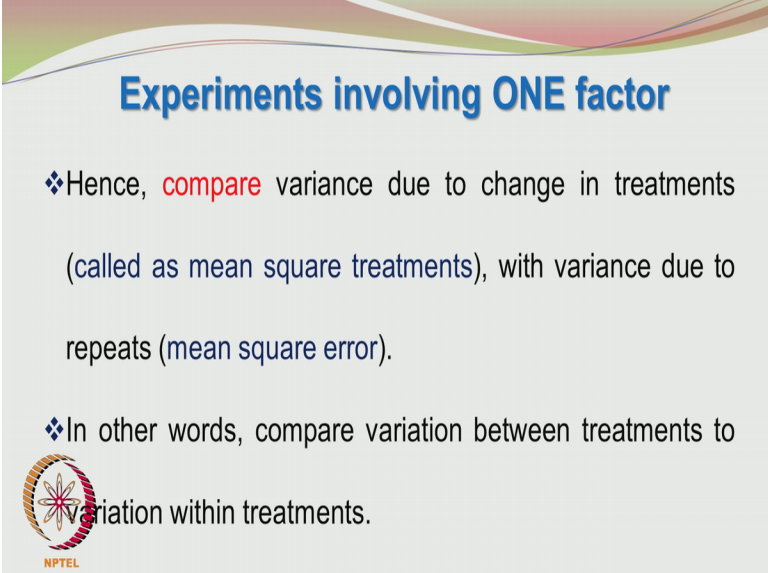
So these are very interesting, and when you look at the T-distribution you have to note that there is a degree of freedom as additional parameters, even for chi-square distribution you have degrees of freedom as additional parameters. When you have F-distribution you have numerator degrees of freedom, and denominator degrees of freedom. So when you have such distributions, you cannot have one single probability chart where you give the appropriate statistic value and then find the probability of the table.

It may require extrapolation or some interpolation, and that is not going to be very accurate. Fortunately, when you look at statistical software or even spreadsheets, they are having powerful statistical functions, and you can find both the probability and also inverse of the probability

using these functions. For example, if you are given the f value, you can find the probability or given the probability you can find the f value, you can find the inverse of the probability to find the f value.


So all these things are possible, so I request you to be familiar with the spreadsheet or statistical software where you can calculate all these probabilities without any ambiguity. Now let us move on to experiments involving one factor, and here we are going to assume that the experimental response is influenced by changing only one variable, all other variables or factors are kept at constant values.

(Refer Slide Time: 42:15)



Experiments involving ONE factor

- ❖ Hence, **compare** variance due to change in treatments (called as mean square treatments), with variance due to repeats (mean square error).
- ❖ In other words, compare variation between treatments to variation within treatments.

 NPTEL


And here we look at mean square treatments and mean square error, we look at the variability caused by changing the level of that particular factor, and we also do repeats and we look at the variability in the repeated runs. We compare the variability across treatment levels with the variability due to repeats, and find the ratio, if the variability across treatment levels are much higher than the variability due to repeats.

Then we can claim that despite the experimental error the factors influencing the outcome of the experiment. So these are very interesting concepts, and since we have only one factor the mathematical analysis is not difficult, and you can easily do the calculations even by having a hand calculator, and doing these calculations with spreadsheet is even easier.

(Refer Slide Time: 43:26)

Analysis of Variance Table (ANOVA)

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{treatments}}$	$a-1$	$MS_{\text{treatments}}$	$MS_{\text{treatments}} / MS_{\text{error}}$
Error	SS_{Error}	$a(n-1)$	MS_{error}	
Total	SS_T	$an-1$		



So we construct the analysis of variance or ANOVA table, where we list out the source of variation, the form of treatments and error, we have the total source of variation then we have the sum of squares, sum of squares of treatments, and sum of squares of errors, we have total sum of squares, the degrees of freedom for the treatments for $a-1$ where a is the number of levels of the factor settings.

For example, if I am looking at the effect of temperature on the reaction yield, then if I have 4 different temperatures, then the number of treatment levels is 4. So the degrees of freedom would be $4-1$ or in general $a-1$, a is the number of independent factor levels or settings. And then you also calculate the sum of squares due to error, let us say that each factor setting you are repeating the experiment n times, then the degrees of freedom associated with the error are $a*n-1$.

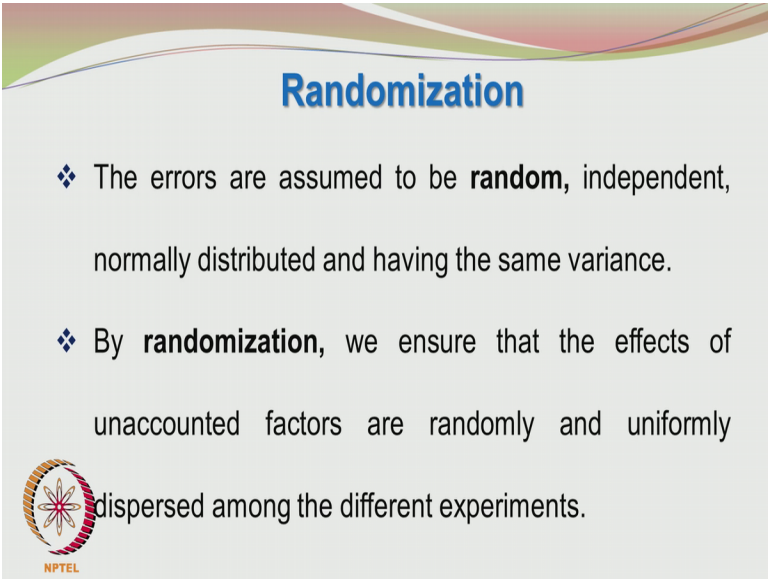
Now you next calculate the mean square, mean square is nothing but the ratio of the sum of squares of treatments by the degrees of freedom for the treatment, so we get mean square treatments. Then we also have the mean square error, which is nothing but the ratio of the sum of squares of the error by the degrees of freedom for the error, so we have mean square error. And you calculate the f value based on the ratio of the mean square treatment to mean square error.

And we see whether this f statistic falls in the critical region, you have a level of significance α and you find out, what is the f_{α} corresponding to the treatment degrees of freedom and error degrees of freedom. The treatment degrees of freedom would appear in the numerator and the error degrees of freedom would be termed as the denominator degrees of freedom. So when you find out the value of f_{α} numerator degrees of freedom, denominator degrees of freedom.

You will have the critical value, you see whether this ratio exceeds the critical value, and if it exceeds then the statistic is lying in the rejection region, and you can reject the null hypothesis. If however, the f statistic is lower than the f_{α} numerator denominator degrees of freedom, then you have to accept the null hypothesis which states that the treatment is not having an effect on the response.


All the variation is only caused by random fluctuations, so it is important for you to state the hypothesis clearly and unambiguously, and then carry out the f test and make the correct conclusion. One important thing to note here is whenever you do experiments you please do as many repeats as possible, so that you have an idea about the experimental error.

(Refer Slide Time: 46:40)



Randomization

- ❖ The errors are assumed to be **random**, independent, normally distributed and having the same variance.
- ❖ By **randomization**, we ensure that the effects of unaccounted factors are randomly and uniformly dispersed among the different experiments.



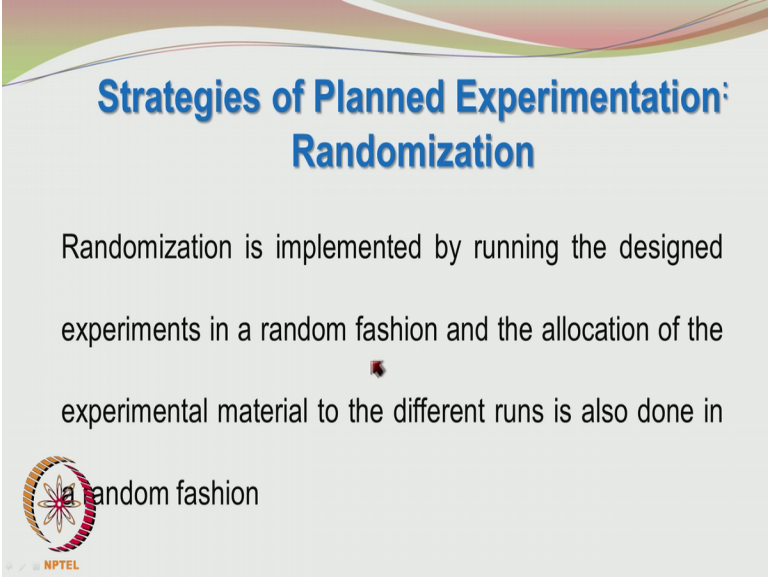
NPTEL

It is important for you to randomize your experiments, so that the sequence is not coming in a standard manner but in an arbitrary or random manner. So this is to ensure that any discrepancies present in the experimental data over and above the variation caused by the main factors are only

due to random factors, and not any systematic factors. So any deviations or any scatter in the data may be attributed only due to random errors, and not due to systematic discrepancies, so we have to do randomization.


And when we do randomization, the effects of the unaccounted factors are randomly and uniformly dispersed among the different experiments.

(Refer Slide Time: 47:33)



**Strategies of Planned Experimentation:
Randomization**

Randomization is implemented by running the designed experiments in a random fashion and the allocation of the experimental material to the different runs is also done in a random fashion

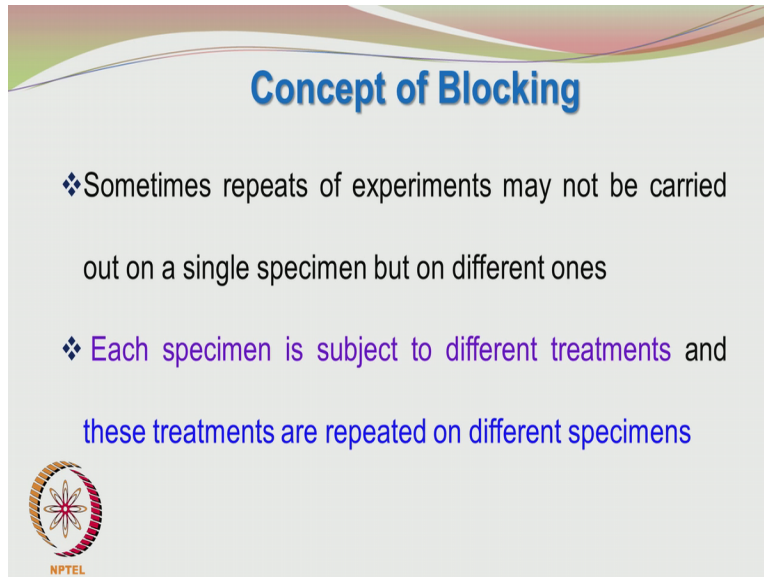
 NPTEL

So the randomization is implemented by running the designed experiments in a random fashion and the allocation of the experimental material or resources to the different runs is also done in a random fashion. You jumble of the sequence of your runs, so that there is no specific pattern. For example, if your experiment involves running at the larger speed of the machine, and it also happens that particular day is very, very hot.

Then you are having the effect of high speed and high temperature from the ambient that may have a particular effect on the variable, so you make it higher than anticipated response, but if you do the experiment in such a way that medium speed and low speeds of the machine operation are also carried out on hot days, then the effect is sort of dispersed among the various settings of your factors.


So this randomization is very important, so that any interference from the external world is sort of distributed to all the runs, and not only to a specific set of runs, and hence they do not seriously affect our process.

(Refer Slide Time: 48:56)



Concept of Blocking

- ❖ Sometimes repeats of experiments may not be carried out on a single specimen but on different ones
- ❖ Each specimen is subject to different treatments and these treatments are repeated on different specimens



NPTEL

An additional issue here is blocking; this we will continue in the next lecture.