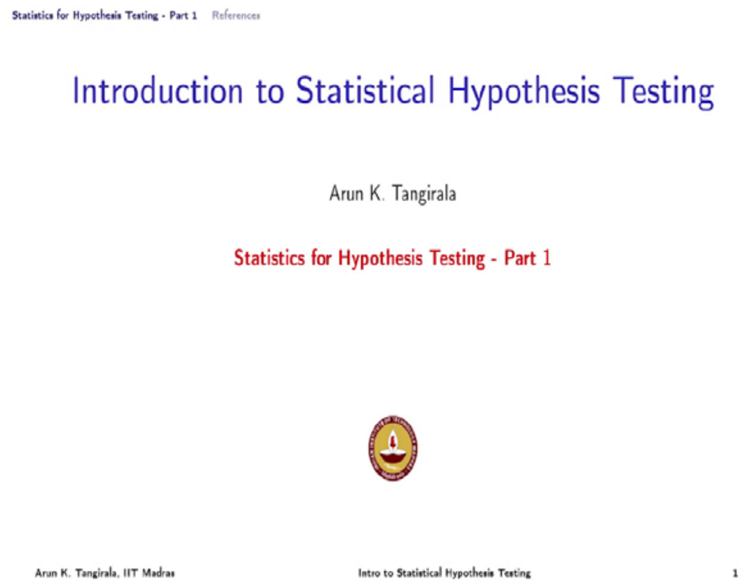
Introduction to Statistical Hypothesis Testing

Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture - 05
Statistics for Hypothesis Testing – Part 1

(Refer Slide Time: 00:09)



Statistics for Hypothesis Testing - Part 1 References

Introduction to Statistical Hypothesis Testing

Arun K. Tangirala

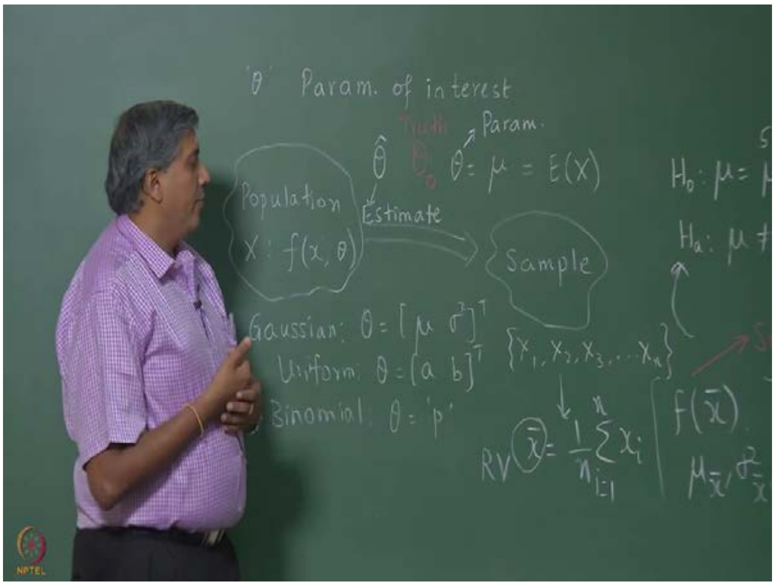
Statistics for Hypothesis Testing - Part 1

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 1

Hello friends, welcome to the lecture on the Statistics for Hypothesis Testing, as a part of the course on Introduction to Statistical Hypothesis Testing. Before I list the objectives of this lecture, it will be a good idea to recap what we have learnt in the previous lectures. Now, if you recall we had a general procedure for hypothesis testing that we went over in the motivation lecture and let me make use of the board to explain what we have done. And, now where we are getting?

So, what we have said in hypothesis testing that is as a general procedure.

(Refer Slide Time: 01:00)


'θ' Param. of interest
Truth
Param.
$\hat{\theta}$ θ_0 $\theta = \mu = E(X)$
Population
$X : f(x, \theta)$
Estimate
Sample
Gaussian: $\theta = [\mu \ \sigma^2]^T$
Uniform: $\theta = [a \ b]^T$
Binomial: $\theta = 'p'$
$\{x_1, x_2, x_3, \dots x_n\}$
RV $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
$f(\bar{x})$
$\mu_{\bar{x}}, \sigma^2_{\bar{x}}$
$H_0: \mu =$
$H_a: \mu \neq$

We said we will first identify the parameter theta that we want to conduct our hypothesis test on, so this is the parameter of interest. So, what is this theta denote? Where does it all come from? Well we said there is a random phenomena which is characterized by a PDF and we said use the term population to denote all possible outcomes of this random experiment or random phenomenon, and this population is characterized by this random variable x. Let us say, this could be our solid a propellant burning rate or may be the diameter of a nylon connector or anything else as well. Associated with this random variable we have either a PDF, if x is continuous value or a P M F Probability Mass Function if x is discrete value. Now, this PDF here is typically characterized by certain parameters theta, if it is a Gaussian distribution for example this theta would be mu and sigma square. For a Gaussian, the parameters of PDF are mu and sigma square that is mean and variance. If it uniform distribution then theta are the range, theta is a vector of the range of values of the distribution. And then, if it is a binomial random variable then you have the P theta is actually the P the probability of success and so on.

So, every PDF that characterizes a random process or a random phenomenon is characterized by these parameters and in basic statistical analysis these are the parameters of interest that is the parameters of the PDF. In a more advanced analysis such as in may be linear regression or in time series modeling this theta would include more than just the parameters or the PDF for example, theta could include model parameters such as the slope or the intercept and so on. However, restrict ourselves to these kinds of analysis to begin with when we go to linear regression we will include more parameter, more elements into this theta. Now, this is the first step in hypothesis testing if you recall identify the parameter of interest. Let us take an example to understand what we have trying to do here, let us take the solid propellant burning rate, example number 2 in the motivation lecture where we were interested in testing whether the average propellant burning rate is 50 centimeters per second. So, we are interested for example in mu, so the theta n is simply mu. What about sigma square, when we are testing hypothesis of this form we may assume that the variance is known or unknown, we will even simplify the problem assume that the variability is known. Now in order to conduct such a hypothesis test what I have to do is of course, ideally what I would like to do is look at entire population and use the definition of mu which is that it is an expected value of x; that is, it is average value of all possible burning rates that is I.

Look at all possible specimens coming from the manufacturing process and then experimentally determine the burning rate take the average clearly that is an impossible task to finish in finite time. Therefore, what we do is we collect a few specimens randomly and then constitute what is known as a sample; that is the basic idea. So, we are moving from the (Refer Time: 05:19) or the population space or the outcomes space to the sample space where the sample consist of a subset of this and what kind of a subset a random subset this that randomness is very important, so that we do not introduce any bias in our eventual decision that we will take based on the sample.

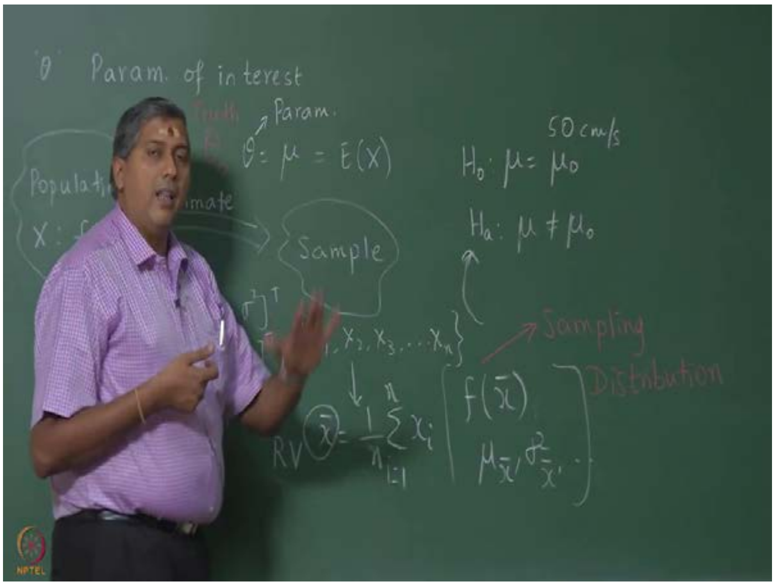
We have spent quite a bit of time in the last few lectures understanding, what is the theta? How are the theta defined? What is the interpretation? What is a PDF? What is the probability mass function and so on; so, mostly we have learned the theory although it is a bit boring but it is a necessary boredom you can say, so that when we move to the sample space this is the practicality, this is theory, here we have let us say n observations of the same random variable denoted by the subscripts, so subscripts here denotes the specimen number, x is the same random variable of interest to make. In the solid propellant burning rate x is the propellant burning rate. Now, my goal in statistical data analysis as per as hypothesis testing is concerned is to take these n observations and then draw inferences of this mu which I do not know through this sample mean that we work with so frequently. The big x are the random variables, the small x are the values that this random variables take. The burning rate that comes out of each specimen is actually an a random variable and I am adding up n such random variables to obtain x bar, why I am doing this so that because I believe that this x bar is an estimate of the truth, which I do not know and hypothesis testing is about making some comments verifying certain or testing some truth statement about mu based on the estimate which is x bar this called a Sample mean.

In general, statistical inferencing is about drawing inferences about the population from the observation. For this specific example of hypothesis testing of mean we may work with this estimate or some other estimate this need not be the only estimate for mu. There are other ways of estimating mu for example, I can use sample median sample mode and so on. But the truth is one that is what should be remembered. So, the entire idea now is we are moving from the population (Refer Time: 08:09) we have learned the theory that helps us understand this statistical description of this population or a probabilistic description of this population to sample space where we are going to work with what are known as statistics.

So we are moving from probability world to statistical world and what is a statistic in general, any function of the random sample that I collect is a statistic that means, a random sample that I collect here it is not specimen the random sample consist of burning rates of those specimens. Whenever I subject it to a mathematical operation and produce a new number then that new number becomes a statistic. Of course, we have to choose statistics based on what we are trying to estimate and the important thing that should be remembered is since each x i is a random variable, x bar also inherits the d n a and therefore is a random variable this is the main things.

So now, in hypothesis testing what we were going to do is use x bar as a means of testing some claim like this.

(Refer Slide Time: 09:16).



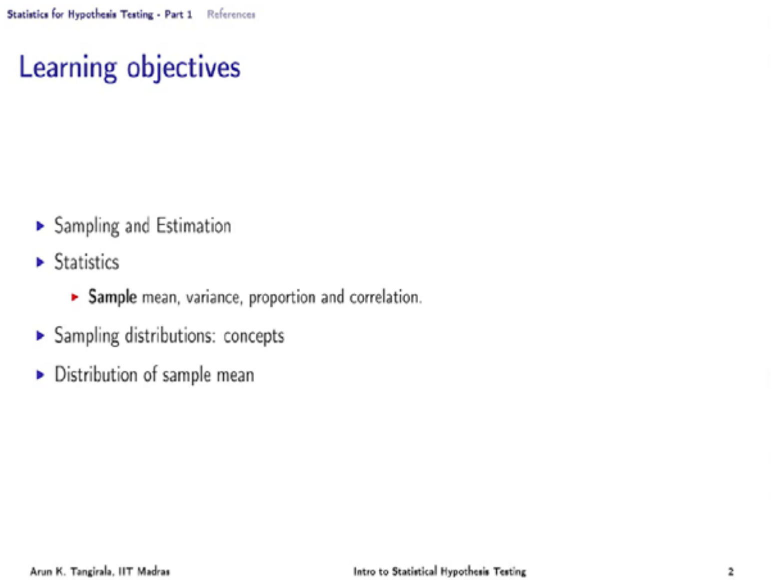
Where the null hypothesis is mu equal something mu naught and the alternate hypothesis for example could be it is naught, so 2 sided tests. In the solid propellant burning rate example mu naught was 50 centimeters per second and of course in the alternate hypothesis as well. So, from probability to statistics for the purpose of hypothesis testing, what we learn in this lecture is we are going to learn briefly about how to sample, but mostly about what is a statistic how do I analyze this a statistic because this is a random variable, it also has it is own PDF that is a very important point to remember. The moment you say something is a random variable it will have it is own PDF and rightfully, it has it is own mean and it has it is own statistics and so on, sorry variance and so on. So, like any other random variable it has it is own PDF, it has it is own mean, it is own variance and may be other moments. What statistical analysis will tell us is given some information about where this x's came from, that is some idea of the population not necessarily fully obviously, if I know fully then I do not have to sample I know everything about the population given partial or no information about the population.

Firstly, how do I construct this PDF and then what can I say about mu and sigma square that is point number 1 and point number 2, very importantly for the course which is hypothesis testing, how do I now use this x bar and this information to carry out the hypothesis test? I am using x bar as a means or as a vehicle for carrying out this test and I will show you through a simple illustration in (Refer Time: 11:36) as to why this piece of information is very important for conducting the hypothesis test. That is a crux of the hypothesis testing.

Simply, constructing an estimate will not allow us to conduct a hypothesis test. We have to go further and construct the PDF somehow either through theory, there is a theory for that and then or through stimulation’s what we call as Monte Carlo stimulation, that is more of an advance stuff. We will mostly work with cases where I can theoretically construct PDF of x bar and then from this I can actually carry out the hypothesis test. At this moment for beginners it may not be as obvious as to, how the PDF of this estimate is responsible or it is going to help us in carrying out this hypothesis testing. And one more piece of information that I want to give is we will generally use theta as the parameter and theta hat as it is estimate or as a statistics. This usage of hat in estimation theory or in statistical analysis is quite common to denote that it is only an estimate and different from truth. Theta is the parameter and theta hat is the estimate and theta naught is a truth which I do not know. So, I do not know the truth and I am going to draw some conclusions about this based on the estimate, but before I do that I will have to construct the distribution of the estimate, this is what we call as the Sampling distribution.

Hopefully, now it gives you a feel of where we were and where we are going to go in this and the next lecture. Once we have understood all of this and a simple example of how to use this to conduct a hypothesis test, then the procedure fortunately is as same for all parameters, that is very nice of this hypothesis testing that it has a same procedure for all parameters in that is why we outline the general procedure. Therefore, what we shall do is initially will go through a simple example may be a few minutes into the lecture, which will give us an idea of why this is important in hypothesis testing. What is this? This is called the Sampling distribution, this f of x bar is called the Sampling distribution or also called the distribution of the estimate either way; we will use the term sampling distribution. First an illustrative example where will just argue intuitively and use some common sense to understand what is the roll in hypothesis testing, and then gradually realize may be a couple of lectures later how exactly we use this in hypothesis testing.

(Refer Slide Time: 14:57)



Statistics for Hypothesis Testing - Part 1 References

Learning objectives

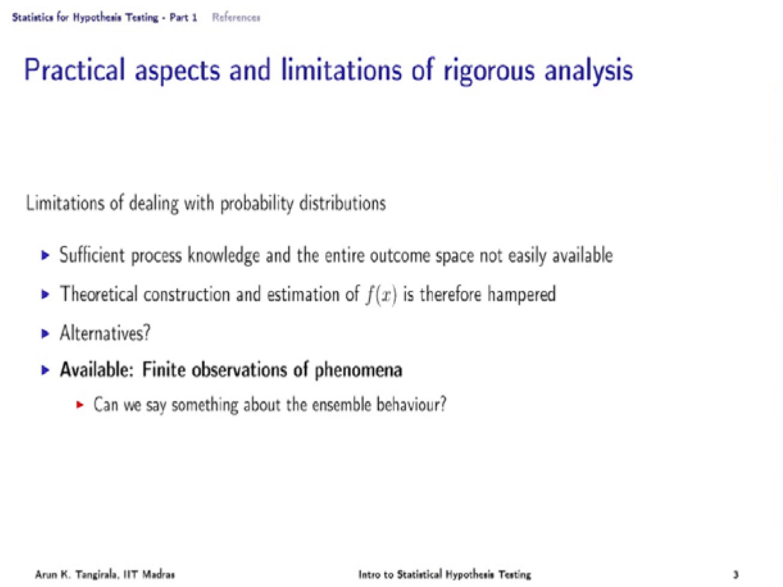
- Sampling and Estimation
- Statistics
  - **Sample** mean, variance, proportion and correlation.
- Sampling distributions: concepts
- Distribution of sample mean

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 2

So, with that prelude let me list the objectives will learn briefly about sampling and estimation I will take this opportunity to briefly talk about estimation because as you have seen in the black board discussion that what is at the core or crux of hypothesis testing is estimation.

Only when we estimate, we can proceed towards hypothesis testing or even for that matter any other data analysis exercise. Once we discuss those then we will spend some time on statistics and also spend a lot of time on sampling distribution concepts. In this lecture we will only talk of distribution of sample mean exactly what I have discussed on the board and in the subsequent lectures we will look at distribution of sample variance, sample proportion and sample correlation.

(Refer Slide Time: 15:54)



Statistics for Hypothesis Testing - Part 1 References

Practical aspects and limitations of rigorous analysis

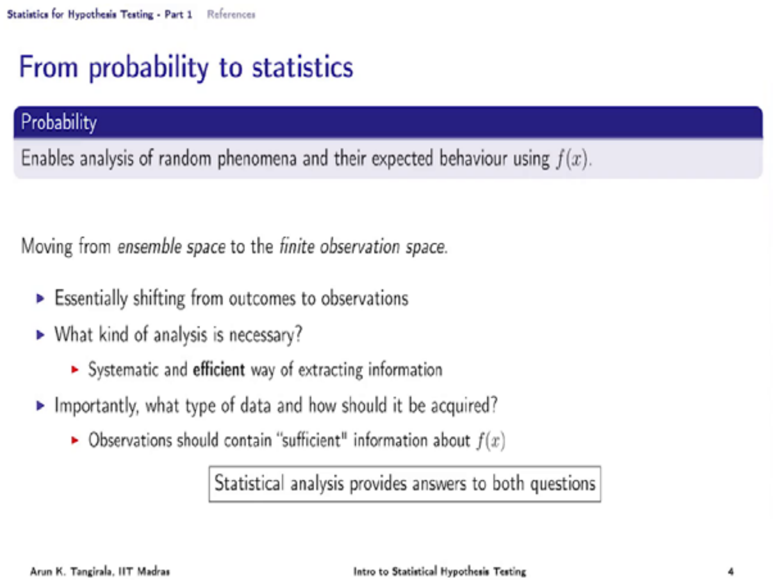
Limitations of dealing with probability distributions

- Sufficient process knowledge and the entire outcome space not easily available
- Theoretical construction and estimation of $f(x)$ is therefore hampered
- Alternatives?
- **Available: Finite observations of phenomena**
  - Can we say something about the ensemble behaviour?

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 3

I have already explained this as to why we are moving from probability to statistics because of practical reasons. There is no way I have access to the entire population, if I have that then there is nothing to worry about in the sense I just get the truth straight away, it is that difficulty that I have, where, which is forcing me to look at sample random subset. So, the main point is we have only finite observation of phenomena and from this finite observation can we say something about the population that is the main question of interest.

(Refer Slide Time: 16:39)



Statistics for Hypothesis Testing - Part 1 References

From probability to statistics

Probability

Enables analysis of random phenomena and their expected behaviour using $f(x)$.

Moving from *ensemble space* to the *finite observation space*.

- Essentially shifting from outcomes to observations
- What kind of analysis is necessary?
  - Systematic and **efficient** way of extracting information
- Importantly, what type of data and how should it be acquired?
  - Observations should contain "sufficient" information about $f(x)$

Statistical analysis provides answers to both questions

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 4

As I said earlier, what we are doing is we are moving from probability to statistic, where we are moving from the anomalous space to the observation space we can say so. And in the course of this statistical analysis, that is analysis of observations and the uncertainties arising as a consequence of the uncertainty in the data that is what all the constitutes statistical analysis. We have 2 important questions to ask, what kind of analysis is necessary? Obviously, that depends on the purpose of statistical data analysis. In this course, the purpose is hypothesis testing, in some other exercise it may be regression; in other exercise, it may be classification and so on. So, what kind of analysis is necessary depends on the application. And then, what kind of data do we actually require? How do we collect data? Are there some guidelines? Are there some rules that I have to follow? Is there a theory to it? Yes, there is a lot of theory to data acquisition in statistical data analysis which is club put together under the banner of sampling. Pretty much like in signal processing what you have a sampling, but here it is not about in signal processing it is all about thinking of what sampling rates should be chosen and so on.

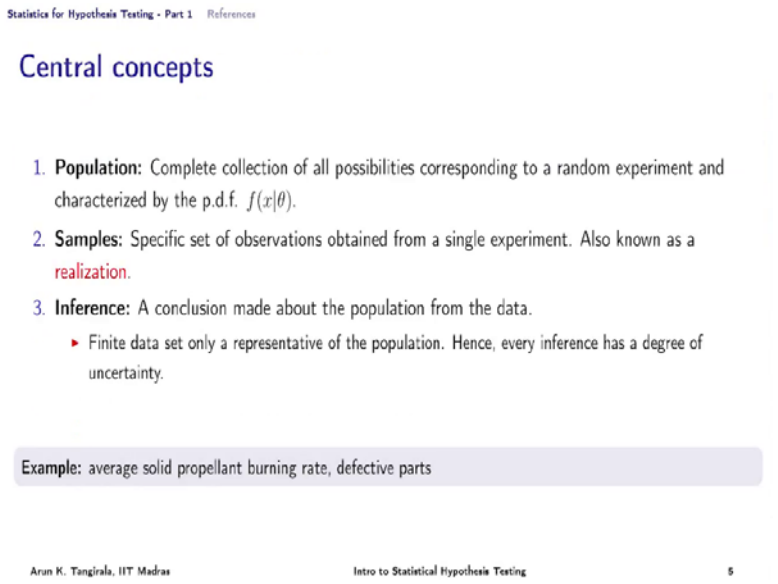
Here, it is not about necessarily sampling rate it is about how I collect the sample mainly within a sample whether the observations should be independent or can be dependent then what is a consequence and all such questions are addressed. And, also very importantly when I collect observations they should have sufficient information about the parameter of interest clearly. Suppose I am looking at let us say I want to know what is the average temperature outside this room and I am going to look at let us say the heights of the population here in the city and from their I want to infer the average temperature outside clearly, there would not be much information about or no information about the average temperature from the height readings of individuals in that city correct.

Then we say that the data is not informative at all. Now, if you turn to the world of statistical data analysis, there is a quantity known as Fishers Information introduced by Fisher which is quite wonderful it characterizes the amount of information that is present in the data with respect to a certain parameter theta. We are not going to talk about that at all. In a fully fledged, let us a time series analysis or statistical data analysis course we will have many opportunities to do so. So, the bottom line is statistical analysis not only helps us in analyzing data that is giving us methods to analyze data, but also tells us how we should collect data and whether the data is informative and so on, clearly because the quality of analysis depends on the data. So, any analyst should be clearly worried about the quality and the information content in the data.

Many a times we may not control over how the data was generated, but knowing the impact of uncertainty in the data on the final analysis is always helpful because, let us say you have a data which is of poor quality or not informative and so on you can straight away say that either upfront that you can say that well the inferences that I am going to draw are going to be highly unreliable and then who ever supplies the data may have to work again repeat the experiment to produce better quality data.

In this course, of course we are not going to worry about any of that very briefly will talk about sampling, but mostly focus on the analysis again in the context of hypothesis testing on. There are 3 central concepts as you have listen to me speaking during the prelude to this lecture, which is that of population, which is the collection of all possibilities of a random experiment and then a statistic or the samples which actually is a subset of the population that I am observing.

(Refer Slide Time: 21:10)



Statistics for Hypothesis Testing - Part 1 References

Central concepts

1. **Population:** Complete collection of all possibilities corresponding to a random experiment and characterized by the p.d.f. $f(x|\theta)$.
2. **Samples:** Specific set of observations obtained from a single experiment. Also known as a realization.
3. **Inference:** A conclusion made about the population from the data.
   - Finite data set only a representative of the population. Hence, every inference has a degree of uncertainty.

Example: average solid propellant burning rate, defective parts

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 5

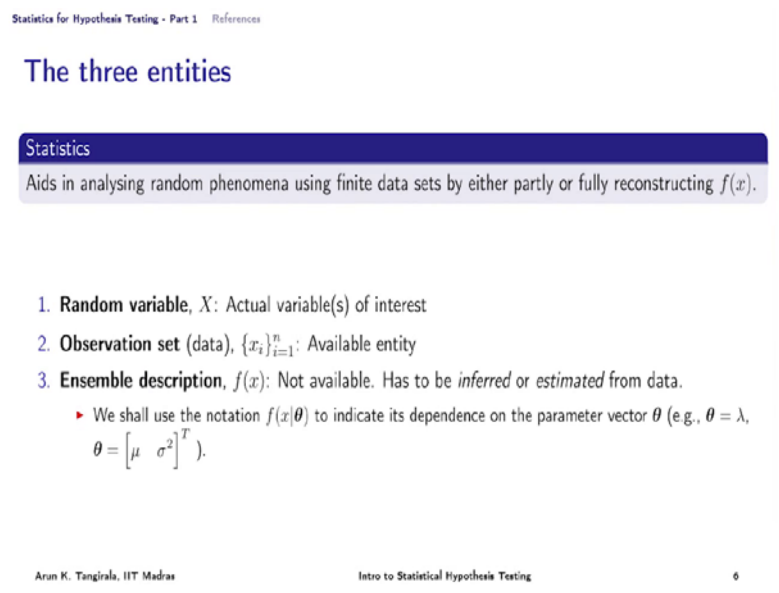
Many a times in time series analysis the single sample is called a Realization, because we use 1 censor to collect small n number of observations and what we have is not the truth, but the truth as perceived by the censor or as realized by the censor. Imagine that many of us (Refer Time: 21:37) there are many who watch movie in a movie theater, imagine that situation where the movie that is being screened is only once truth is only one, but each person in the audience perceive a particular scene entire movie in a different way.

If you were to ask the opinions or the review ratings they would vary widely perhaps across the audience, know to movie watchers may give you exactly the same review. So, what is happening is that the truth is one, but what the person is reporting is his or her perception. The censor is actually watching some phenomenon observing, some phenomenon, sensing some phenomenon and giving you the truth plus it is own characteristic which has the randomness and that is a why a different censor can give you a different reading. We are actually working with what is known as one data record or one realization.

This term realization is mostly used in signal analysis not so much in statistical data. Then we have inference, of course this is the concept that we have been speaking about drawing conclusions about the population based on the observed data. And, what we have is a finite data set and we have to live with that that is the point. The main point to remember is, since I have not seen the entire population and I am only going to see a finite subset of the population whatever inferences I draw from this finite set will always be in error. For example, yes in the earlier discussion on the blackboard the sample mean that I am using to estimate mean will never be identical to the truth, as long as the number of observations is finite no way you can get actually the truth. Unless the observations that you have obtain is the entire population which is not necessarily the case for us.

So, that is a point to remember x bar is always going to be different from mu or theta hat is always going to be different from theta naught.

(Refer Slide Time: 23:51)



Statistics for Hypothesis Testing - Part 1 References

The three entities

Statistics

Aids in analysing random phenomena using finite data sets by either partly or fully reconstructing $f(x)$.

1. **Random variable**, X: Actual variable(s) of interest
2. **Observation set** (data), $\{x_i\}_{i=1}^{n}$: Available entity
3. **Ensemble description**, $f(x)$: Not available. Has to be *inferred* or *estimated* from data.
    - We shall use the notation $f(x|\boldsymbol{\theta})$ to indicate its dependence on the parameter vector $\boldsymbol{\theta}$ (e.g., $\boldsymbol{\theta} = \lambda$, $\boldsymbol{\theta} = \begin{bmatrix} \mu & \sigma^2 \end{bmatrix}^T$).

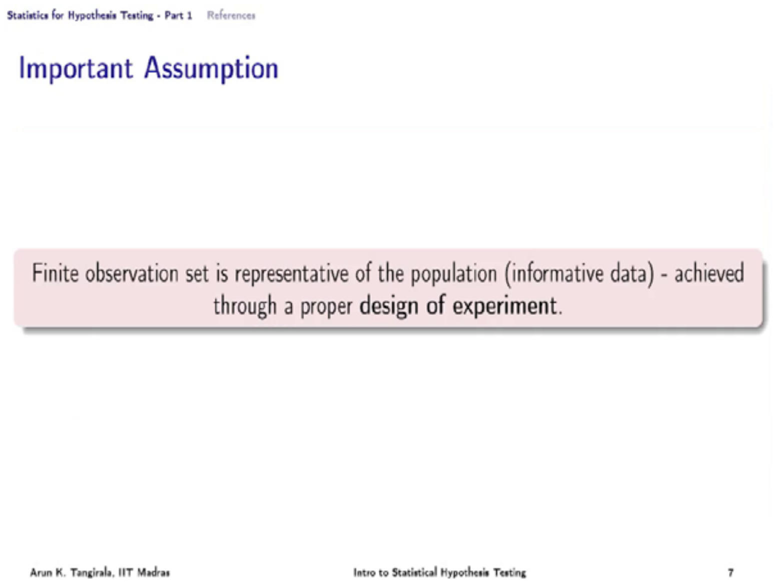
Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 6

Apart from the 3 concepts, we have 3 entities in statistical data analysis.

One is the random variable, that is an entity which is the main object of interest main variable of interest based on which I am actually formulating the entire problem. In the solid propellant burning rate the burning rate is a random variable or in the proportion of defective items in a manufacturing process the state of the part that I am analyzing whether it is defective or non-defective that is a random variable and so on.

So, that is the number 1 entity that is going to participate in the statistical data analysis. And then of course, is your observation set from which you are going to draw conclusions about the (Refer Time: 24:40). There is a truth which I know partially or I do not know, if there is a random variable which I know exactly what it is stand for and then there is a data which I have collected. These are the 3 ingredients that participate in the statistical data analysis.

(Refer Slide Time: 24:58)



Statistics for Hypothesis Testing - Part 1 References

Important Assumption

Finite observation set is representative of the population (informative data) - achieved through a proper **design of experiment**.

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 7

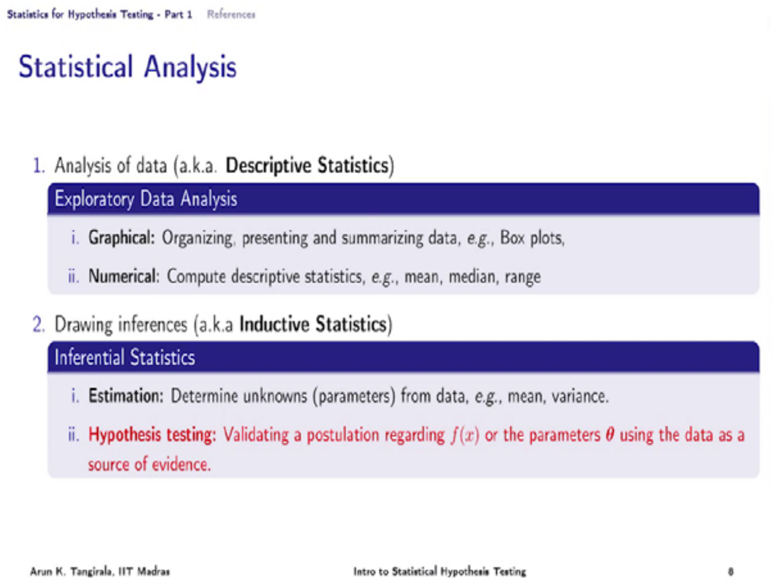
And there is an implicit assumption whether we state it or not I am sure it is common sense, that when we collect this is finite observation set we assume that it is representative of the process.

Now, all of that I am saying is not just valid for hypothesis testing, it is also valid for in general statistical inferences. Slowly we will drive ourselves towards hypothesis testing, but these are the generic concepts that have to be learnt regardless. Now, this is an important assumption that the finite observation set is representative of the true population. As an example, suppose I want to find out what is the most visited park or I want know if I take a particular park, I want to know, what is the average number of people visiting this particular park in a day? Obviously, the number of people visiting a particular park will vary across days and it is quite unpredictable, so I can think of the number of people visiting particular a park as a random variable and I am interested in knowing, what is the average number of people visiting that particular park? Now, I will not have access to the entire population therefore, I will have to randomly sample some individuals who come to this park and somehow carry out my experiment to figure out the average number of people.

Now, if I am doing my experiment properly I will actually select the proper set of individuals, but if I am not so careful about the experiment I may end up actually selecting people only who come daily to the park. And as a result, if all the individuals in my sample are people who come daily only somehow because by virtue of sampling then I may end up with a wrong answer. On the other hand, I should just randomly sample the number of the individuals who come to the park and ask them, how many times they have gone in a month? And, that will probably constitute a better sample.

That is what we mean by proper or a careful sampling and also mean that the finite observation truly represents the population. So, I should not confine myself systematically or knowingly towards a subset of the population, it should just be randomly done and that is idea of random sampling.

(Refer Slide Time: 27:42).



Statistics for Hypothesis Testing - Part 1 References

Statistical Analysis

1. Analysis of data (a.k.a. **Descriptive Statistics**)

   Exploratory Data Analysis

   i. **Graphical:** Organizing, presenting and summarizing data, *e.g.*, Box plots,

   ii. **Numerical:** Compute descriptive statistics, *e.g.*, mean, median, range

2. Drawing inferences (a.k.a **Inductive Statistics**)

   Inferential Statistics

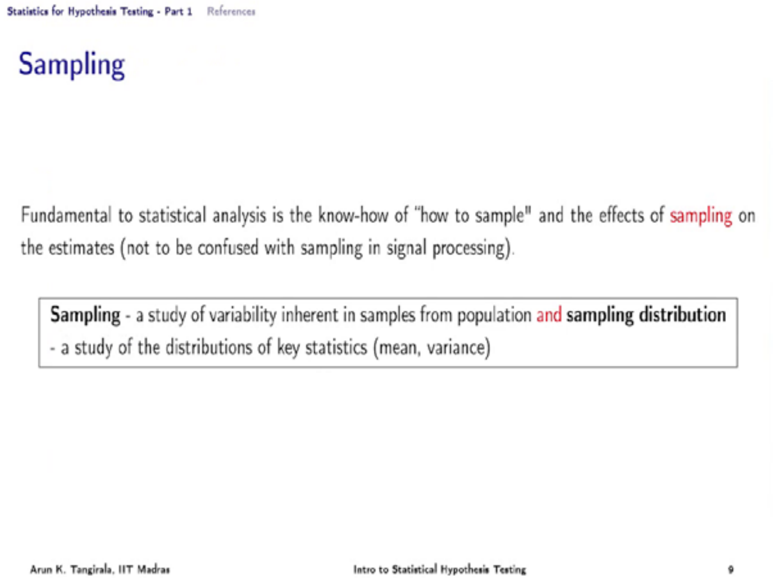
   i. **Estimation:** Determine unknowns (parameters) from data, *e.g.*, mean, variance.

   ii. **Hypothesis testing:** Validating a postulation regarding $f(x)$ or the parameters $\boldsymbol{\theta}$ using the data as a source of evidence.

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 8

And, now let us say once we carry out the sampling.

(Refer Slide Time: 27:48)



Statistics for Hypothesis Testing - Part 1 References

Sampling

Fundamental to statistical analysis is the know-how of "how to sample" and the effects of sampling on the estimates (not to be confused with sampling in signal processing).

Sampling - a study of variability inherent in samples from population and **sampling distribution** - a study of the distributions of key statistics (mean, variance)

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 9

Let me actually quickly talk about sampling and then come back to move on to analysis. So, sampling is this act of obtaining subset of data from a population and a random sample is that which is obtained in such a way that there is no systematic or known bias in your observations. The other way of saying that is all the observations are independent of each other which mean a joint PDF of the observations is going to be a product of the marginal PDF of each observation. Now, let us talk about statistical analysis. So, I have now samples with me and now I want to proceed towards analysis, as I said I am just taking this opportunity to talk about some general things about statistical analysis and then very quickly we will move on to inferential analysis for hypothesis testing.

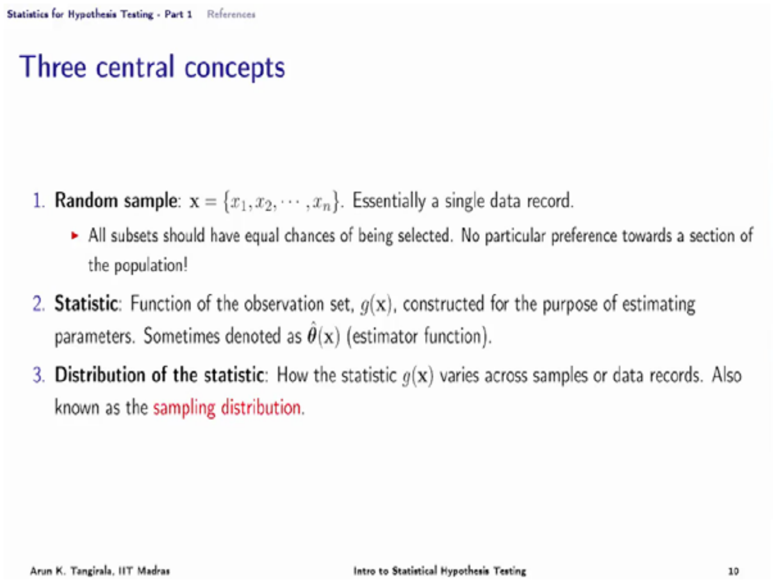
Typically, the first step in data analysis even in hypothesis test you should be doing this in fact, is to visualize the data spend some time with the data do not simply take the data and load it into some software package and then report the outcome. Then you are not really the user is not needed I can write programs that will automate all of this. The users roll is in interpreting the results in the context of the process from it is data has been obtained and also collaborate the results qualitatively with the visual analysis and see if everything is meaningful and so on. There is still a quite a bit of user intervention that is required and the first intervention that is required is graphical analysis where we first look at the data in the form of displays such as scatter plots, box plots or pie charts and so on. So, there are number of ways in which you can graphically represent the data and the kind of display that you would choose obviously depends on the data and the application, but standard plots include scatter plots or bar plots or box plots where they immediately, they tell you what is the range or what is if there is a trend or if there is an oscillatory component or if there is an out layer, where is the center located and so on.

There is a lot of analysis that we can make through a visual inspection and we should never undermine the roll of that visual inspection. Once we have done that then we can move on to numbers. As an example of the importance or the power of the human brain in visual analysis let me give you a simple example where we looking at an oscillatory wave in the context of pattern recognition, if I am looking at a sign wave then with the human eye let us say there is only single frequency instantly we recognize that there is an oscillatory behavior and bit more careful analysis will also help us reveal the frequency of the oscillation.

Now, to do this mathematically and with a help of a computer 1 has to convert the signal into a fourier transform into the fourier domain through the help of fourier transform, compute the spectrum search for the peak and then locate the frequency at which this peak occurs. There is so much of processing that is required. We do not know at the moment how are brains are able to detect so easily, of course a visual inspection has it is own limitation, but there is a lot of information qualitative information and sometimes quantitative that we can gather through visual examination of the data which is otherwise quite difficult through a mathematical or a statistical analysis. Once we are through with identifying all the out layers making amendments for the data and so on then we generate numbers which will tell us more about the data, what we could not gather visually. What we mean by numbers is computing mean, computing median, mode and so on. Those are the preliminary kinds of steps that we normally take and we should also do that in hypothesis testing. It is not that simply I take the number generate the estimate or the statistics then generate the PDF the distribution and then apply it plug it into the hypothesis testing procedure and that is not how we do it.

Once we have done with that and then we move on to drawing conclusions where we move into the world of inferential statistics, this is where now I am going to estimate certain parameters and draw some conclusions about the population or the population parameters. And, in particular we are interested in hypothesis testing which is concerned with validating some postulates about the population based on the estimates that I have updated.

(Refer Slide Time: 32:55).



Statistics for Hypothesis Testing - Part 1 References

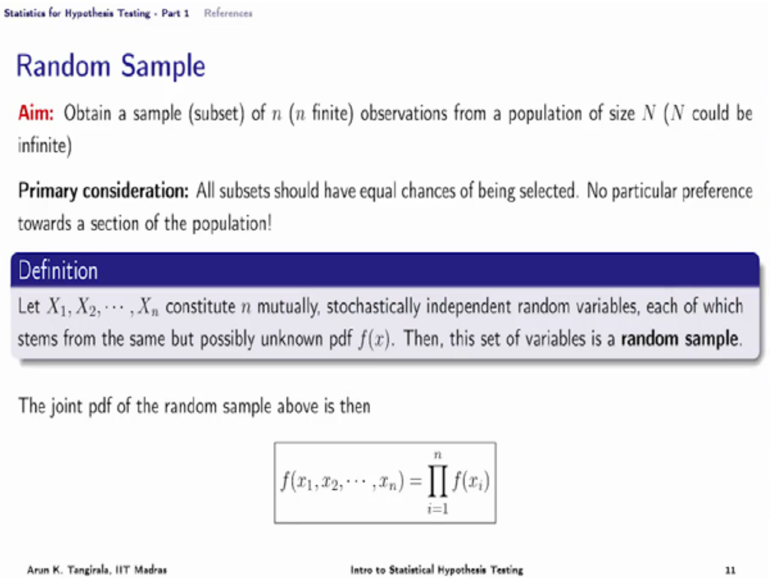
Three central concepts

1. **Random sample**: $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$. Essentially a single data record.
   - All subsets should have equal chances of being selected. No particular preference towards a section of the population!
2. **Statistic**: Function of the observation set, $g(\mathbf{x})$, constructed for the purpose of estimating parameters. Sometimes denoted as $\hat{\boldsymbol{\theta}}(\mathbf{x})$ (estimator function).
3. **Distribution of the statistic**: How the statistic $g(\mathbf{x})$ varies across samples or data records. Also known as the sampling distribution.

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 10

And again, associated with this statistical analysis as we know is a random sample there is an statistics. So, what do I need for statistical analysis, I need sample that is an observation set a data set and then I need a statistic that means an x bar for example, which it is a statistic or any other estimate that I am going to compute and then the distribution as I said earlier you have the observation data in the mean estimation example from which you compute x bar and we also compute f of x bar. These are the 3 things that participate in inferential statistics not the full statistical data analysis, but inferential statistics or you can say hypothesis testing.

(Refer Slide Time: 33:48)



Statistics for Hypothesis Testing - Part 1 References

Random Sample

Aim: Obtain a sample (subset) of n (n finite) observations from a population of size N (N could be infinite)

Primary consideration: All subsets should have equal chances of being selected. No particular preference towards a section of the population!

Definition

Let $X_1, X_2, \cdots, X_n$ constitute n mutually, stochastically independent random variables, each of which stems from the same but possibly unknown pdf $f(x)$. Then, this set of variables is a **random sample**.

The joint pdf of the random sample above is then

$$\boxed{f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i)}$$

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 11

What is a statistic? What is a formal definition of a statistic? The formal definition of a statistic is, as I had explained earlier is that it is some mathematical function of the data that you have obtained. The g is the 1 that we use to denote this function and the output is theta hat, strictly speaking I should not have a theta hat there I should have some other random variable, but we will assume that the statistics are being calculated for the purpose of estimation. Now, the reason I say this is there is a settle difference between what a statistic is? And what an estimate is? A statistic is some general function of the observations that I am probably computing or calculating for the sheer joy of it I do not know right, there may not be a great purpose to computing the statistics.

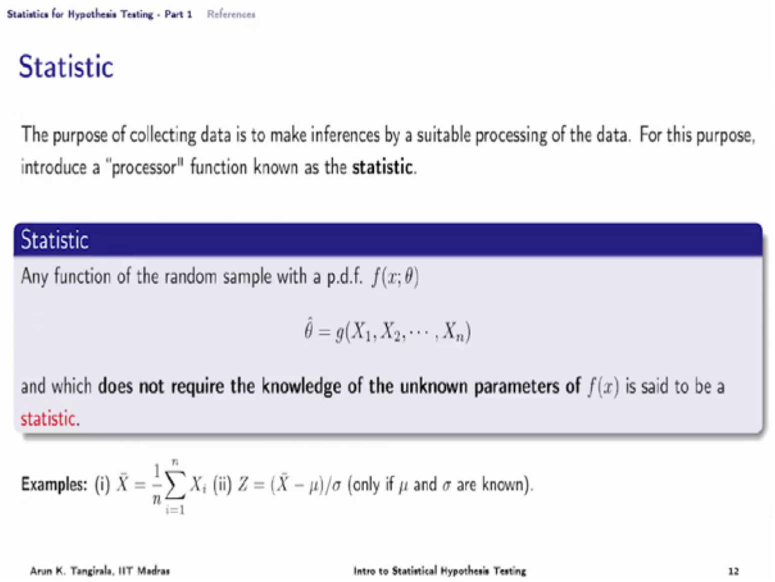
Whereas, with an estimate there is a definitive purpose, there is a goal in a mind I want, I am estimating; I am subjecting the data through this function g, so as to estimate some parameter. That is the prime difference between a statistic and an estimate, but we will not observe such differences in this course we will use the term estimate and statistic interchangeably because we are going to use statistics only for hypothesis testing not for the sheer fun of it. So, will you denote theta hat, statistic by theta hat and remember we have to keep telling this to ourselves that theta hat is also a random variable in it is own right because it is being computed from a (Refer Time: 35:32) of several random variables a collection of random variables, all are the uncertainties actually will creep into theta hat as well. Remember that there is yet another requirement for the statistic which is that it does not require the knowledge of the unknown parameters of f for example, if I take sample mean then to compute a sample mean all I need is a data I do not need to anything more.

Alright, so let us look at these two examples at the bottom I have x bar which is a sample mean one over n sigma x i that is a random variable, x bar is a random variable it is a statistic because on the right hand side I know everything I do not need any additional knowledge. Whereas, in the second example I have a new random variable constructed from x bar which we do routinely leave, in fact we should get use to this in hypothesis testing of means where I am constructing what is known as a standardized sample mean. X bar minus mu by sigma let us say; now this z is called as a statistic if and only if I know mu and sigma. What are this mu and sigma? This mu is the mean of the PDF which is generating the data for me and sigma is let us say the standard deviation of the PDF or sigma could be the standard deviation of x bar as well which I do not know, will show later on that the standard deviation of x bar can be calculated knowing the standard deviation of the PDF that generating the data and the sample size.

That sigma can be thought of as sigma x by route n, in any case the z is called a statistic only when mu and sigma are known. Whereas, x bar is a statistic always because a right-hand side is not a function of any parameter of the PDF. That is the settle point that also 1 has to remember in calling something as a statistic. And, I think you should now relate this word statistic to the steps that we have outlined in the general procedure in the motivation lecture, we said identify the parameter of interest collect data and then compute the statistic.

What is that test statistic? The test statistic is the nothing but the statistic that you are going to use for the test which is being computed from the data.

(Refer Slide Time: 38:08)



Statistics for Hypothesis Testing - Part 1 References

Statistic

The purpose of collecting data is to make inferences by a suitable processing of the data. For this purpose, introduce a "processor" function known as the **statistic**.

Statistic

Any function of the random sample with a p.d.f. $f(x;\theta)$

$$\hat{\theta} = g(X_1, X_2, \cdots, X_n)$$

and which **does not require the knowledge of the unknown parameters of** $f(x)$ is said to be a statistic.

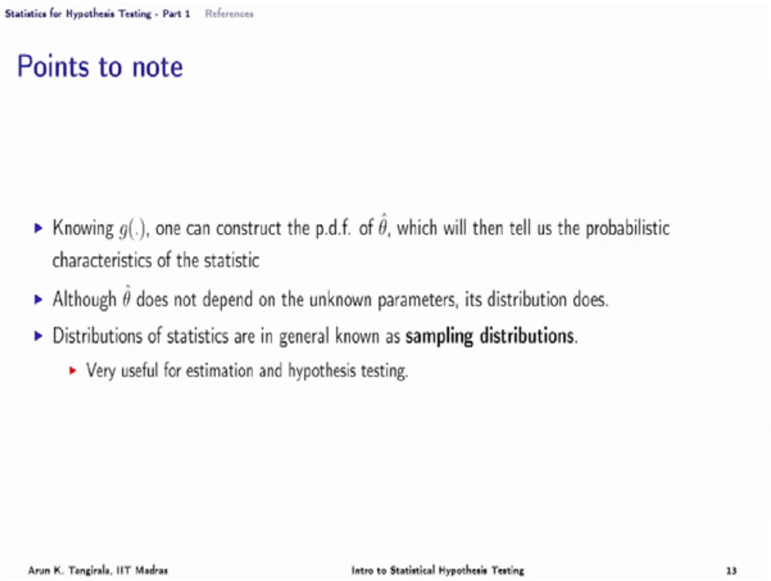
Examples: (i) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (ii) $Z = (\bar{X} - \mu)/\sigma$ (only if μ and σ are known).

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 12

That is what is test statistic is. So, 3 few important points to note again something that have been repeatedly saying, knowing g typically we know the function or the formula you can say loosely that we are going use to compute the statistic or the estimate. We can construct the PDF of theta, now this is not such an easy task in many situations in some situations it is quite easy depending on the function g for example, if g is linear like in the sample mean case it is easy to construct PDF of theta hat. On the other hand, if g is non-linear like we see may be in some other complicated estimation exercise may be not in this course then it becomes difficult to derive the PDF of theta hat. Essentially, the nature of g determines how easy it is going to be for me to determine how the randomness in x will propagate to randomness in theta hat, if it is linear it is very easy I can use central limit theorem and so on.

If it is non-linear then very rarely may be in 1 or 2 circumstances I can actually find out, but by and large I have to resort what are known as Monte Carlo stimulations to figure out the PDF of theta head, but I need that for hypothesis testing I need f of theta hat. Fortunately, in this course we are going to work with statistics for which the PDF of theta hat is already known in theory I do not have to really turn to simulations for any help that is the good news. And, also again the technical term as I said earlier on the board, this distribution of statistics are known as sampling distribution which are very useful in estimation and hypothesis testing.

(Refer Slide Time: 40:00)



Statistics for Hypothesis Testing - Part 1 References

Points to note

- Knowing $g(.)$, one can construct the p.d.f. of $\hat{\theta}$, which will then tell us the probabilistic characteristics of the statistic
- Although $\hat{\theta}$ does not depend on the unknown parameters, its distribution does.
- Distributions of statistics are in general known as **sampling distributions**.
  - Very useful for estimation and hypothesis testing.

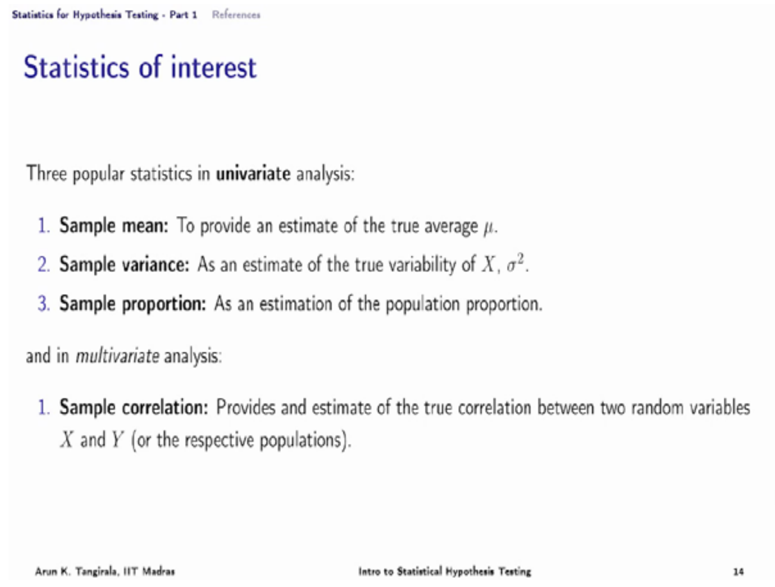
Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 13

Now, we are going down further we are slowly narrowing down to our problems of interest which is that of hypothesis testing of mean, variance, proportion and correlation. Therefore, the statistics of interest for us are sample mean, sample variance, sample proportion and in bivariate analysis sample correlation. Now, having said this I should also tell you that the hypothesis test are for the mean, variance and proportion and correlation 1 of these, it does not mean that I have to use only these statistics I can use any other statistics that will help me test those hypothesis. As an example, suppose I want to test the hypothesis for the solid propellant burning rate case which is the hypothesis test on the mean, I do not have to use sample mean as a statistic I can use sample median also right, I can use any other estimator of mean. The only reason why we choose sample mean is of course, it is simple and a lot of times it is good and it has some nice properties, it is PDF can be determined easily and so on.

However, if you know that the data is prone to out layers. Your data contains can be contaminated with the few out layers then sample mean can actually end up being a wrong choice, may be sample median is a better choice because sample mean is not a robust statistic. If you look at the theory of statistics there are different classes of estimators, robust estimators, unbiased estimators and so on. Robustness has got to do with how sensitive this statistic is to out layers in data, in which case maybe it is best to use sample median. But if you are sure there are no out layers then it is better to work with sample mean because it will give you the least error estimate it is called efficient estimate. So, the bottom line is that the hypothesis test is whatever it may be mean, variance and so on I can use any relevant statistic to test the hypothesis. The choice of the statistic depends on a few considerations, what kind of data? Whether the data contains out layers or not? And equally importantly how easy it is for me to obtain the PDF of that statistic alright. Now, before we really plunge into the sampling distributions of the sample mean and so on, let me take this opportunity to also tell you that statistics as I said the word statistic is more or less equivalent to estimate with some settle differences that I pointed out.

And likewise, the inferential statistics is analogous to estimation or is a part, in fact, of the estimation. Estimation theory is a very broad feel that is applicable to all branches of engineering, medicine, social sciences and so on.

(Refer Slide Time: 43:19).



Statistics for Hypothesis Testing - Part 1 References

Statistics of interest

Three popular statistics in **univariate** analysis:

1. **Sample mean:** To provide an estimate of the true average μ.
2. **Sample variance:** As an estimate of the true variability of X, σ^2.
3. **Sample proportion:** As an estimation of the population proportion.

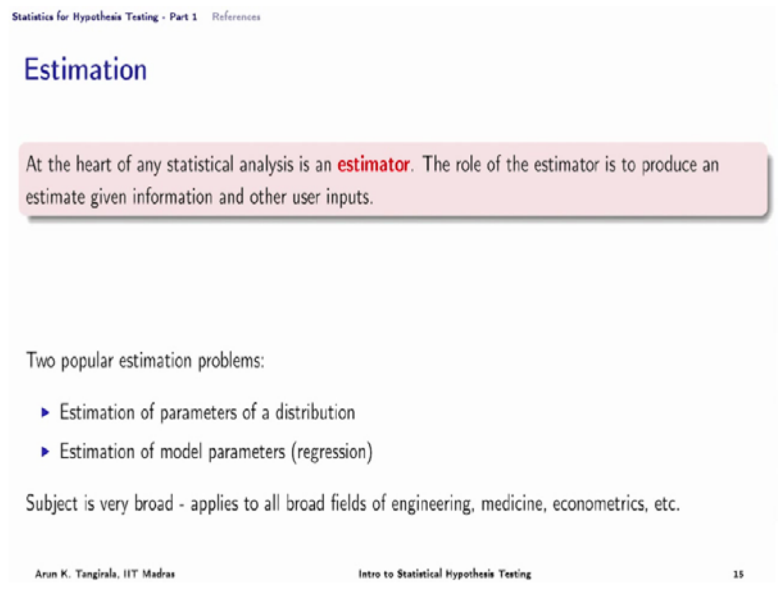
and in *multivariate* analysis:

1. **Sample correlation:** Provides and estimate of the true correlation between two random variables X and Y (or the respective populations).

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 14

And, at the core of any statistical analysis is estimation as we have been discussing. It is important to be also familiar with term estimation and we may be interested in estimation theory in estimating parameters of a distribution or model parameter or may be signals and state space modeling may state estimation and so on. In this course, we are only obviously going to be worried about estimation of parameters of distribution and estimation of model parameters in linear regression because all our hypothesis test are in those context only. The estimation is a very broad theory.

(Refer Slide Time: 43:56).



Statistics for Hypothesis Testing - Part 1 References

Estimation

At the heart of any statistical analysis is an **estimator**. The role of the estimator is to produce an estimate given information and other user inputs.

Two popular estimation problems:

- Estimation of parameters of a distribution
- Estimation of model parameters (regression)

Subject is very broad - applies to all broad fields of engineering, medicine, econometrics, etc.

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 15

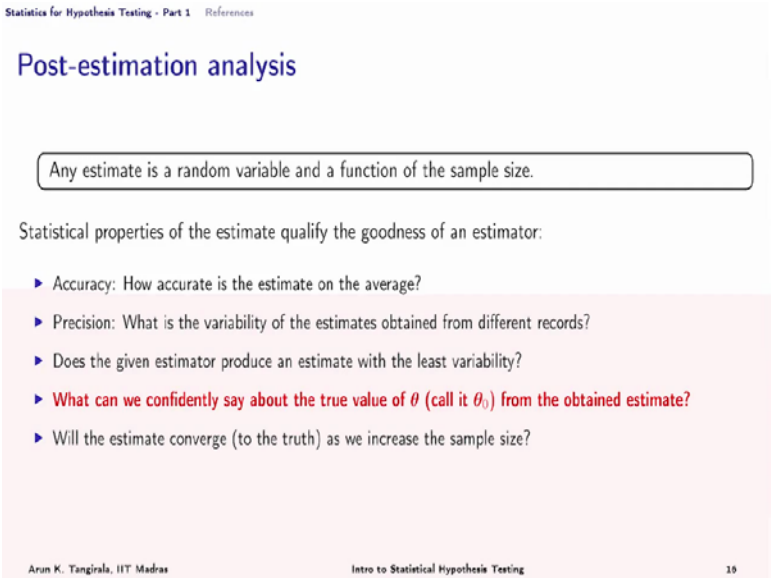
And again, a point that I have been saying, it is very important to actually tell this to ourselves that any estimate that I construct from data is going to be a random variable because it is a function of the sample size right and of course the randomness in the data. So, in this context there are several questions that 1 asks in estimation, the common question that asks is, how good is the estimate? Right; in the estimation of mean, if I use sample mean I would like to know, how good is the sample mean as an estimator? What do you mean by good, there are 2 things that quantify goodness, but both are talking about the closeness of the estimate to the truth. 1 measure is accuracy which looks at the averages of the estimates. Now, here is where I just want to tell you as to why an estimate is a random variable another way of looking at it, we have been saying because it is a function of the random sample it is a random variable that is all right.

But, if I want to understand truly, what is this randomness business in an estimate? Let us go through this very simple thought process. I collect data, that means I have 1 data record I performed 1 experiment where I have generated n observations and from where I can construct 1 estimate. I can set another experiment I repeat the experiment under the same conditions, I will get a new data record the values of which will be different from the values that I got in the first experiment. Now, from the second data record I construct another estimate that is I have x bar from first experiment for example, x bar from second experiment and likewise think of these as a thought process where I am conduction infinite experiments, number of experiments I am repeating this again and again and again then for each data record I have an estimate.

Everything held the same I am getting different values of x bar, that is the mark of a random variable right. Remember, we defined random phenomenon as everything a whole constant, but outcomes are going to be different. So, I have held everything experiment is (Refer Time: 46:14) same everything is sample size is same, I am collecting the same number of observations, same censor yet the outcome that is of interest which is an estimate is changing from experiment to experiment to experiment therefore, it qualifies to be a random variable and therefore it has it is own PDF. Let me actually show you shortly an illustrative example highlighting all of this, but let finish this first. So, for just now we said per experiment I have an estimate, suppose I were to take the average of all such estimates from all those experiment and if that average hits a true value then we say it is an accurate or an unbiased estimate right.

On the other hand, I am worried about the variability of the estimate from experiment to experiment which is characterized by the variance of theta hat and the technical term for that is precision, we keep hearing this terms high precision instruments and so on. There it is about sensing instruments that are repeatedly used to measure the same thing I hold a process the same I use a censor repeatedly to see if he gives me the same reading, take a thermometer in your home keep measuring the body temperature at the same location around the same time just repeatedly and see if you it gives me the same readings. But of course, there remember it as a resolution that means it cannot show perhaps beyond a first decimal. You may think it is giving you the same reading, but if you at actually go into the third-fourth decimal and so on you will see even the second decimal it may give you a different reading. So, there is going to be variability in the readings there, here we are talking of variability in the estimate they are one and the same almost.

(Refer Slide Time: 48:03).



Statistics for Hypothesis Testing - Part 1 References

Post-estimation analysis

Any estimate is a random variable and a function of the sample size.

Statistical properties of the estimate qualify the goodness of an estimator:

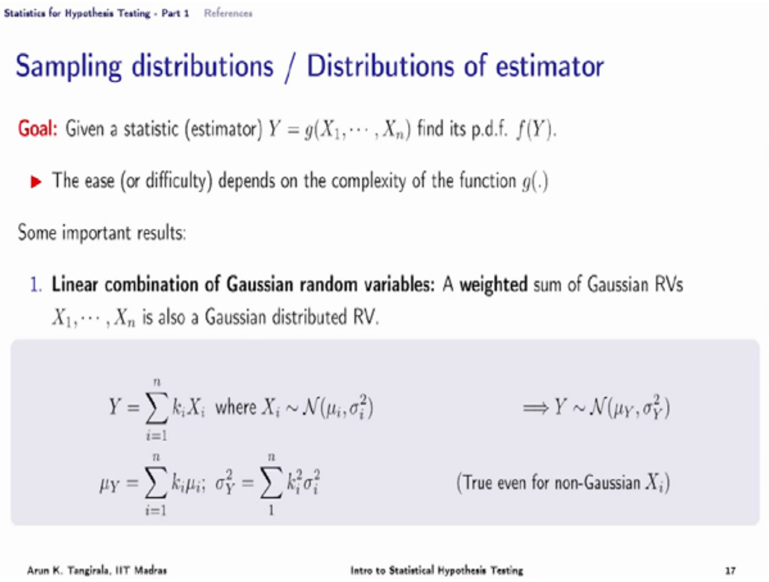
- Accuracy: How accurate is the estimate on the average?
- Precision: What is the variability of the estimates obtained from different records?
- Does the given estimator produce an estimate with the least variability?
- What can we confidently say about the true value of θ (call it θ_0) from the obtained estimate?
- Will the estimate converge (to the truth) as we increase the sample size?

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 19

We say here that the estimate is more and more precise as the variability in the estimates across the experiments decreases, as the variability becomes lower and lower the estimates become more and more precise. Which is more important to us in estimation typically precision is a lot more important than accuracy, both are important so given a chance we would like to have an accurate and precise estimator, but if you have to sacrifice one for the other normally one sacrifices accuracy for precision. Because of a simple reason if one sacrifice accuracy sorry, precision for accuracy then that means I am ok with having high variability of estimates. What does it mean? That means, I am in obtaining an estimate of let us say 6.4 in for the sample mean from one experimental record.

Whereas, another experimental record would have generated may be 9.7; that is a very high variability. What does it mean, that means that the re-estimate that I am constructing from an experiment is not reliable. The precision is a measure of reliability, from 1 experiment alone I may not be able to draw meaningful conclusions or reliable sorry not meaningful but reliable conclusion it is going to be large error in the estimate across experiment I do not want that and therefore, precision is a lot more important than accuracy never the less both are important given a chance. And, then there are other kinds of questions, the most important question in estimation that is of interest to us is what can I say confidently about the true process from the estimate is exactly what we have been saying right from the beginning of this lecture.

(Refer Slide Time: 50:05)



Statistics for Hypothesis Testing - Part 1 References

Sampling distributions / Distributions of estimator

Goal: Given a statistic (estimator) $Y = g(X_1, \cdots, X_n)$ find its p.d.f. $f(Y)$.

- The ease (or difficulty) depends on the complexity of the function $g(.)$

Some important results:

1. **Linear combination of Gaussian random variables:** A **weighted** sum of Gaussian RVs $X_1, \cdots, X_n$ is also a Gaussian distributed RV.

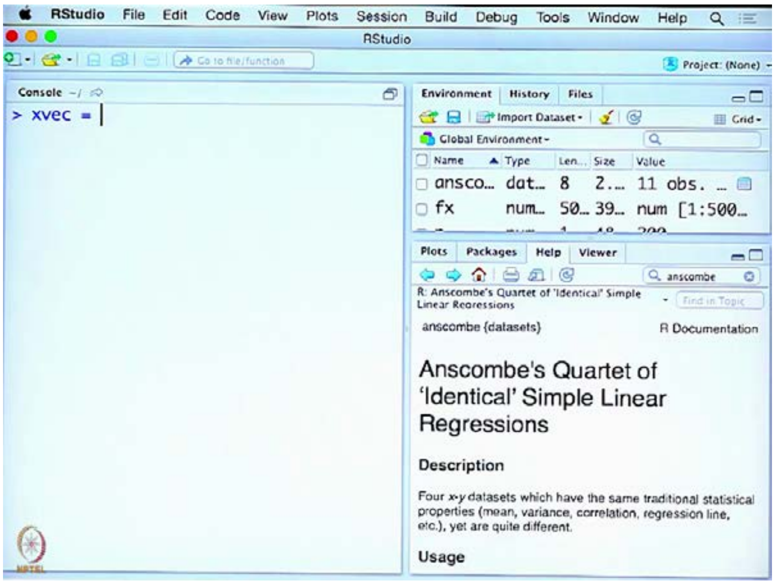
$$Y = \sum_{i=1}^{n} k_i X_i \text{ where } X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \qquad \Longrightarrow Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

$$\mu_Y = \sum_{i=1}^{n} k_i \mu_i; \ \sigma_Y^2 = \sum_{1}^{n} k_i^2 \sigma_i^2 \qquad \text{(True even for non-Gaussian } X_i\text{)}$$

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 17

And, before we move on to sampling distributions that is learning what is the distribution of sample means, sample variance and so on, we may want to look at this general results that will help us in deriving the sampling distribution. Before we move on to learning the sample distributions, it is probably better to get a feel of what this sampling distribution business is all about. We have had a lot of discussion theory and so on, so it will be nice break to move into the practical illustration in a software that we are using which is the r and let me show you a through an illustration what sampling distribution is all about.

(Refer Slide Time: 50:53)



RStudio File Edit Code View Plots Session Build Debug Tools Window Help
RStudio
Go to file/function
Project: (None)
Console ~/
Environment History Files
Import Dataset
Grid
Global Environment
Plots Packages Help Viewer
anscombe
Find in Topic

```
> xvec = 
```

Name	Type	Len...	Size	Value
ansco...	dat...	8	2....	11 obs. ...
fx	num...	50...	39...	num [1:500...

R: Anscombe's Quartet of 'Identical' Simple Linear Regressions

anscombe {datasets} R Documentation

Anscombe's Quartet of 'Identical' Simple Linear Regressions

Description

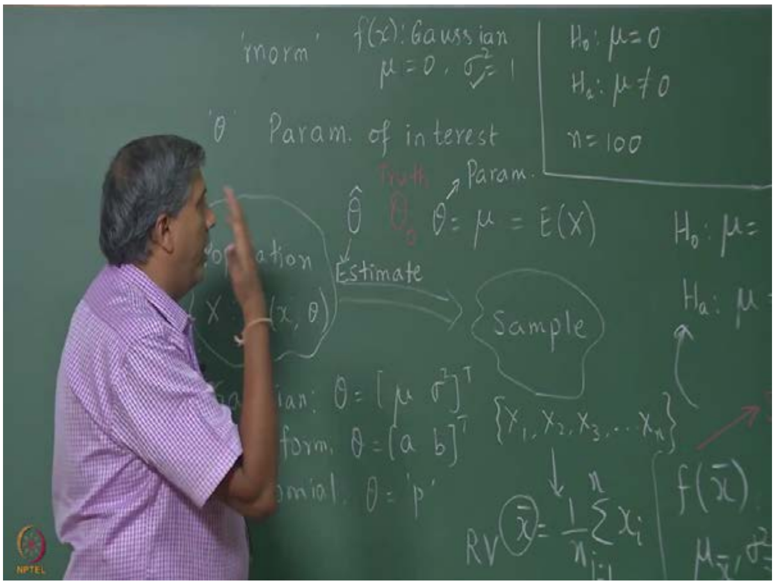
Four *x-y* datasets which have the same traditional statistical properties (mean, variance, correlation, regression line, etc.), yet are quite different.

Usage

Here, I have the r studio in front of me which basically is the d u y and through this we are going to understand what is the concept of a sampling distribution and then get back to the theory.

Let us say, I am going back to the same example of estimating the mean and we want to try and understand what is a sampling distribution of mean or sample mean. The mean of the population is a truth that is a deterministic quantity; sample mean is a random variable. Let us do it this way, let us actually generate a random sample from a Gaussian distributed population. The problem setting is as follows, if I take rnorm which is the random number generator that get me numbers randomly from a Gaussian distribution of mean 0 and standard deviation 1, that is what is the rnorm routine in r. Likewise you have many other in software packages you have similar random number generators. Like for example, in mat lab you have r and m. Now, r claims that the rnorm will get you random numbers from a Gaussian distribution with mean 0 and variance 1. Suppose, I want to actually test this claim made by this particular software; let me write that clearly on the board.

(Refer Slide Time: 52:29)



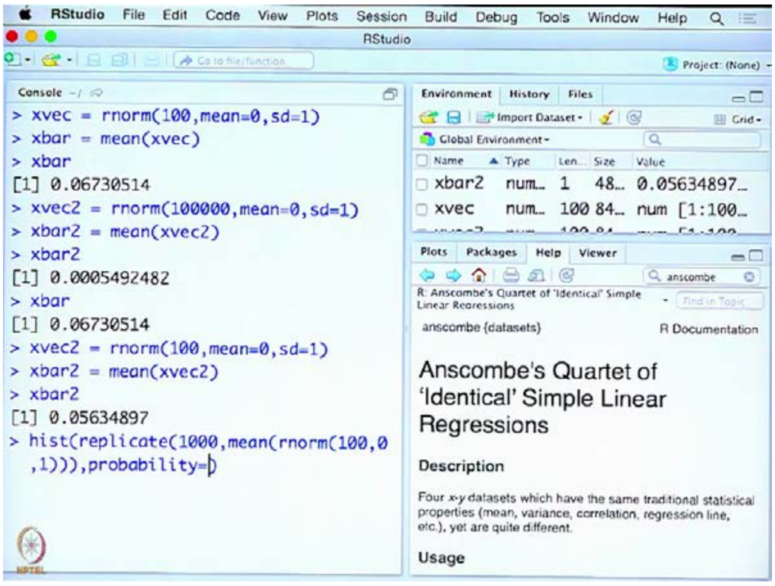
Outline this portion here, what r is actually claiming is that the true mean of the population the Gaussian population generated from which rnorm is sampling is 0. So, r says there is a population here with f of x is Gaussian for rnorm with mean 0 and standard deviation 1, we will write this as variance being 1. This is what when I use rnorm in r is going to sample numbers from this population. Suppose I want to test this claim that truly rnorm getting numbers from this population which means 0.

I will assume that this is ok, this is correct sigma square is 1 I have no contention about that, but only contention that I have is mu is 0. Now I can formulate simple hypothesis test which is null hypothesis is that mean is 0; that is in the absence of any information the default claim made by r holds because I am not doing anything so whatever claim they have they owe the right to take the claim. Once I decide to test their claim, then I formulate this (Refer Time: 54:01) alternate hypothesis no, that is not true.

Now, the only way I can test is, is to sample because the actual possible numbers that you can get from this Gaussian population for all practical purposes is very, very large you can think of it is infinite. So, what we will do is to test this claim we are going to randomly sample 100 observations from this Gaussian population. We are going to choose 100 observations and see if I can claim this. Of course, that is the eventual goal that we want to get to (Refer Time: 54:37) the intermediate step is to obtain us a sampling distribution. We will first illustrate what is meant by sampling distribution and then talk about how that is useful in hypothesis testing.

If you understand this example in it is full clarity, then you have more or less understood the crux of or the philosophy of hypothesis testing. Whether it is mu or some other parameter the procedure is the same. So, now let us randomly sample 100 observations from this Gaussian distribution and syntax for that would be rnorm of 100.

(Refer Slide Time: 55:15)



```
> xvec = rnorm(100,mean=0,sd=1)
> xbar = mean(xvec)
> xbar
[1] 0.06730514
> xvec2 = rnorm(100000,mean=0,sd=1)
> xbar2 = mean(xvec2)
> xbar2
[1] 0.0005492482
> xbar
[1] 0.06730514
> xvec2 = rnorm(100,mean=0,sd=1)
> xbar2 = mean(xvec2)
> xbar2
[1] 0.05634897
> hist(replicate(1000,mean(rnorm(100,0
  ,1))),probability=)
```

Name	Type	Len...	Size	Value
xbar2	num...	1	48...	0.05634897...
xvec	num...	100	84...	num [1:100...

R: Anscombe's Quartet of 'Identical' Simple Linear Regressions

anscombe {datasets} R Documentation

Anscombe's Quartet of 'Identical' Simple Linear Regressions

Description

Four *x-y* datasets which have the same traditional statistical properties (mean, variance, correlation, regression line, etc.), yet are quite different.

Usage

Of course, you can specify to make sure that mean is 0 and variance is 1 in r we specify standard deviation. So, we say this and now I have this vector of 100 observations in x (Refer Time: 55:39). Obviously, if I want to verify if it is Gaussian and so on typically I plot a histogram and so on, we are not interested in that at this moment. What we want to know is; what if I compute the sample mean from these 100 observations? What is the sampling distribution? How do I compute sample mean?

That is very simple I use, so mean computes a sample mean for you and we have stored that in x bar. Let us see what x bar is alright. So, it is this value obviously, we do not expect x bar to be 0 because we have only taken 100 observations, you may think may be if I take 10000 or 100000 observations I may get 0 may be let us do that. Let us take another set of data where I have 100000 observations. Very good, so I have done that now let us compute the sample mean from this fresh data.

So, what is x bar 2, it is very small but it is not 0 and this is going to be the case you may increase a number of observations by 1 more factor of 10, you will get a lower x bar, but it will never hit 0. However, what probably we can see is as I increase a number of observations in my data, in my sample or what we say as I increase the sample size the estimated mean which is x bar seems to be converging to the truth and this what the law of large number essentially tells you that the averages will converts to the truth as you increase the sample size, we will not go in that direction. Right now, what we want to know is, what is meant by sampling distribution of this x bar? So, let us get back to the 100 observation case, what will do is we have already x bar from 1 set of 100 observations.

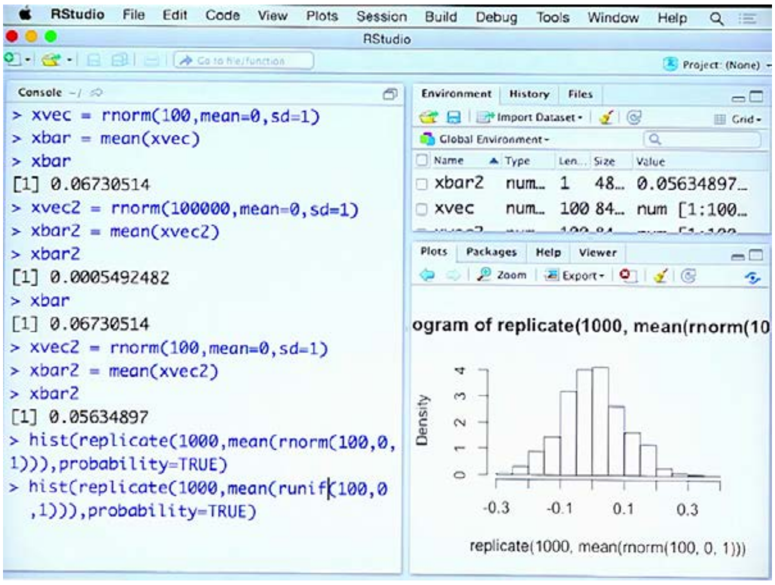
Let us generate another 100 observations, now we will do this and store this in xvec2 and ask, what is the value? Alright, so this is a different value that is exactly what we what I said earlier. Now, this instance here the first instance where we generated is my first experiment, like my first experiment where I have generated 100 observations. And, this instance of generating another sample of the same size is like repeating the experiment same sample size and x bar and x bar 2 are the sample means obtain from these 2 respective experiments and obviously they are different in values, because now I have a different set of 100 values and if I keep repeating this I will get different sets of data and different sets of x bar.

Now, that itself is an indication that it is a random variable. I am repeating the experiment same sample size I get different values. How do I now get the sampling distribution, that is distribution of x bar let us say here. One way is to repeatedly do this may be put this in a recursive operation like for loop and may repeat this perhaps 10000 times. Then collect all the sample means from those respective 10000 replicates of experiments and then plot a histogram because histogram of anything of a collection of data, at least of a random sample sorry will give us a decent idea of the probability density function, in this case occur continuous value random variable. We can actually write this recursive operation put it in a for loop in each iteration of the for loop we generate data compute x bar and store the x bar from each such operation and then plot a histogram this is a sequence of steps we can do.

But, that sounds a lot laborious in r there is a nice way of doing this in a single shot. We will do all of that using this, they will say hist of replicate, right. What does this replicate do? It repeats a experiment. What do we want to be repeated? First we have to specify how many times you want to repeat it let say I want to repeat the experiment 1000 times, more number of times I repeat better will be my estimate of the distribution of x bar. So, what I am going to repeat? Repeat the calculation of x bar. How many observations I want? 100 observations is what I am looking at, correct. And of course, I can specify what is the standard mean and standard deviation; to be clear that what I am doing is correct. Let us say I want the histogram to plot the probability plot just for the sake of it because then I can relate the probabilities very easily; very nice.

So, here I have this plot where what I have on the y axis is the density function, straight away we see that this x bar has a Gaussian distribution right. You can straight away say by looking at the histogram that it has a Gaussian distribution. If there are doubts on this then either we can increase pretty much we can increase the number of replicates or perhaps number of observations. In fact, we will learn in the lecture that theoretically x bar follows a Gaussian distribution regardless of the sample size when the samples fall out of Gaussian distribution. Here, the sample fall out are coming from a Gaussian distribution family, we have 100 not samples it is observations and 100 observations is what I have taken. So, I am linearly combining 100 such observations and theory tells us that when I combine linearly n Gaussian distributed random variables regardless of what is n the resulting variable is also going to follow a Gaussian distribution. That is a property of a Gaussian distributed random variable.

(Refer Slide Time: 62:42)

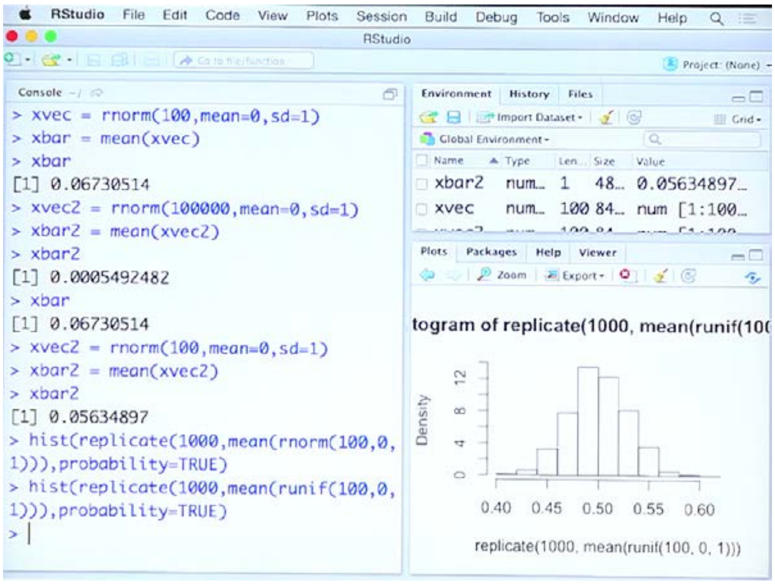

RStudio File Edit Code View Plots Session Build Debug Tools Window Help
RStudio
Go to file/function
Project: (None)
Console ~/
Environment History Files
Import Dataset
Grid
Global Environment
Plots Packages Help Viewer
Zoom Export
ogram of replicate(1000, mean(rnorm(10
Density
0 1 2 3 4
-0.3 -0.1 0.1 0.3
replicate(1000, mean(rnorm(100, 0, 1)))

```
> xvec = rnorm(100,mean=0,sd=1)
> xbar = mean(xvec)
> xbar
[1] 0.06730514
> xvec2 = rnorm(100000,mean=0,sd=1)
> xbar2 = mean(xvec2)
> xbar2
[1] 0.0005492482
> xbar
[1] 0.06730514
> xvec2 = rnorm(100,mean=0,sd=1)
> xbar2 = mean(xvec2)
> xbar2
[1] 0.05634897
> hist(replicate(1000,mean(rnorm(100,0,
1))),probability=TRUE)
> hist(replicate(1000,mean(runif(100,0
  ,1))),probability=TRUE)
```

Name	Type	Len...	Size	Value
xbar2	num...	1	48...	0.05634897...
xvec	num...	100	84...	num [1:100...

So, x bar is actually following a Gaussian distribution. I can test this in many different ways also, for example; I may wonder whether this is true only for this case or it is true for several other cases, so I leave to you to play around with different means and different standard deviations and so on. But observe in each of those 2 things, 1 that x bar follows a Gaussian distribution, 2 look at where the x bar is centered around that is a PDF, sorry PDF of x bar is centered around. On the y axis you have f of x and on the x axis you have the count that is the range of values of x bar. So, for example, many values of x bar in this interval are that is a lot of probability of finding x bar let me put in this way, the probability of finding x bar in this range here within the vicinity of let us say minus 0.1 to 0.1 is quite high. Whereas, the probability of obtaining an x bar beyond minus 0.3 or plus or minus 0.3 is very low right and let me also show you something that whether the samples the observations fall out of a Gaussian distribution or not it is still the x bar will still follow a Gaussian distribution. All you have to do is go and change here, the f of x that is the population distribution to let us say uniform distribution. The point I am trying to make here is whether the observations fall out of a Gaussian distribution or they fall out of some non-Gaussian distribution such as a uniform 1 x bar will still follow a Gaussian distribution.

(Refer Slide Time: 64:46)



```
> xvec = rnorm(100,mean=0,sd=1)
> xbar = mean(xvec)
> xbar
[1] 0.06730514
> xvec2 = rnorm(100000,mean=0,sd=1)
> xbar2 = mean(xvec2)
> xbar2
[1] 0.0005492482
> xbar
[1] 0.06730514
> xvec2 = rnorm(100,mean=0,sd=1)
> xbar2 = mean(xvec2)
> xbar2
[1] 0.05634897
> hist(replicate(1000,mean(rnorm(100,0,
1))),probability=TRUE)
> hist(replicate(1000,mean(runif(100,0,
1))),probability=TRUE)
> |
```

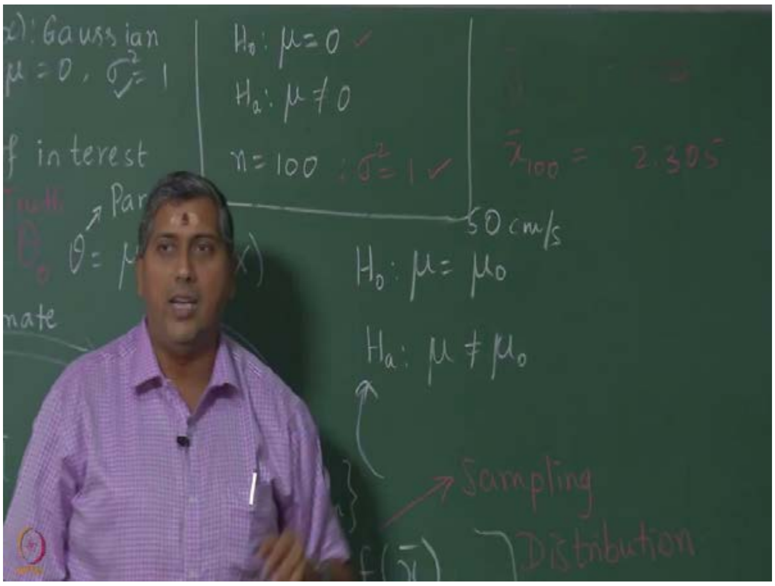
Name	Type	Len...	Size	Value
xbar2	num...	1	48...	0.05634897...
xvec	num...	100	84...	num [1:100...

You can see things have not changed right sorry (Refer Time: 64:55). So, now there is something that you should notice as in between in this plot and the previous plot and let us go back to the previous plot.

The PDF is a Gaussian and is centered around what value the origin, that means the mean of x bar is 0 which happens to be the mean of x also. So, x bar has it is own mean, x has it is own mean. Now turn to the next case where we have generated data from a uniform distribution, again f of x bar is Gaussian and is centered around what point 5 when I changed from rnorm to runif I said I specified this parameter 0 and 1 that means, I am sampling from a uniform distribution within an interval 0 to 1. What is a average of a uniform distribution recall the examples that we went through, the average is a plus b by 2 that is the mean is a plus b by 2 that means, a true minus 0.5. And you can see once again f of x bar or is centered around 0.5 which is the true mean. So, in all of this what we are seeing is that the average of x bar seems to be the same as the x bar the true mean of the population

As we said earlier, that means this is an unbiased estimator of the true mean. Now coming back to the hypothesis testing and we will close the discussion there. Let us go back to the Gaussian case r claims that the rnorm samples from a Gaussian distribution with mean 0 and 1. Alright, now I have to test this hypothesis what I am going to do I am going to do the same thing but not replicate here, I am only in practice I will have only 1 data record, so I will obtain only 1 x bar.

(Refer Slide Time: 67:02).



Now, suppose from one data record, let us say I obtain x bar, so x bar is collected from 100 observations I will indicate this at the bottom of here, it turns out that it is minus 0.15 this is what I get from 100 observations. Alright, now from this obviously x bar is different from the truth and we know it is going to be different from the truth. There are many ways now using common sense I can understand what is hypothesis testing, 1 way of looking at it is well I was suppose to get 0, but I got minus 0.15 should I consider this to be significantly different from the truth. That is 1 way of looking at it, whether the difference between the observed value and the truth is significant. But will pose it in a different way to just shorten the discussion will assume that this to be the truth will say let us say this is the truth.

Let us say whatever r is claiming is correct. Then we ask if this is true and if this is the case that is I have n equals 100 and I am also given that this is correct. I know this is correct this is what I have collected so there is no uncertainty about this and this is what I am going to test. Now, I shall ask the question, if holding all of these things is this a probable how highly or lowly probable this value is? Of course, we cannot ask exactly this but what is the probability for example, of the sample mean taking on values within the vicinity of this. In other words, we are asking if all of this holds good, the null hypothesis holds good, the number of observations being fixed and the variance being known is whatever I have observed a high probable value or a low probable value, less likely or more likely. As an example, suppose I say I worked very hard, a student null hypothesis, student worked very hard.

Alright, but it turns out that the on the exam the performance is very poor. Then the teacher, let us say would want to a test a student's claim there is a probability that even is a student has worked hard the student can end up with poor performance because the paper was actually very, very difficult and probably ask questions that never fell within the preview of the subject. There is a probability, but assume that is very low probable case then the poor performance is not so commiserate with the claim being made by the students that students work very hard as a simple example. So, then what would you conclude you say most likely it is probably that the students did not work very hard something like that.

So, likewise here, if this is true and if the x bar given all the rest of this 2 be true if this turns out be a low probable value then most likely this null hypothesis is not correct. If you turn to the PDF you can see that minus 0.15 is not very unlikely value, but what is likely and what is unlikely is something that we have to decide up front which we call as a critical value. And that depends on the error that I am willing to tolerate. Remember there are going to be 2 different kinds of error, 1 that this null hypothesis is true and I end up rejecting it, the other being this not being true and I still do not reject it. So, 1 is called type 1 error and other is called the type 2 error. Depending on the error that I incur I will fix a critical value. If you look at the probability plot there minus 0.5 is not such a improbable value.

But on the other hand, in some other weird experiment for the same process there may exist 1 realization, it is possible that you when the base (Refer Time: 71:42) and so on. You may end up with some sample such that this turns out be minus 2.305 or plus 2.305. Suppose, this is only realization that you have and got you this x bar; this is no longer valid, this is a realization that you has such that your x bar works out to be 2.3. Obtaining such a realization takes a lot of bad luck.

But let us say you really obtain that and then you obtain x bar this way, now you said this are the only realization that I have. X bar is a fairly reliable estimate that is where the precision of the estimate is important. We say I can rely on the estimator; the estimator is not that bad. Now, I ask given this to be the true how probable is this value. See the probably is extremely low, this is a very rare chance so it is more likely that this is wrong and you end up rejecting the null hypothesis. But all of that depends on what kind of realization you end up with. So, that is what is a crux of hypothesis testing that you will normally hold the null hypothesis being true and then test and that is why the null hypothesis are typically of equality types right. If this was a null hypothesis and then it would have been very difficult to compute the probabilities of x bar or any estimate that is point number 1.

Point number 2; any hypothesis testing will have an error because we are looking at probabilities. And number 3, you should also see and we will see that later that this result of rejecting or not rejecting not only depends on the realization, but depends on these factors. If I change this 100 to let us say 5, then again things will change and so on. But we will see more of this in the next part of the lecture where will talk of sampling distributions of mean and then followed by sampling distribution of variance and so on.

So, we will take a break at this point and then return to the theory in a short while.

Thanks.