

Introduction to Research
Prof. Arun K. Tangirala
Department of Metallurgical and Materials Engineering
Indian Institute of Technology, Madras

Lecture – 14 (2A)
Data Analysis

Hello. Welcome to the course on Introduction to Research; in particular, the lecture on Data Analysis. In the previous lectures, you have learnt what is meant by research. We have tried to give you an overview of research, and of course, we have also gone through a lecture on literature review. In this lecture, we will specifically look at what is data analysis, the types of analysis, and the different types of data that one encounters in data analysis, and of course, a systematic procedure.

(Refer Slide Time: 00:50)


Data Analysis

Objectives

To learn the following:

- ▶ What is data analysis?
- ▶ Types of analyses
- ▶ Different types of data in analysis
- ▶ Systematic procedure

With two hands-on examples in R® (a popular software for data analysis) ...

NPTEL

Arun K. Tangirala, IIT MadrasIntroduction to Research2

We will end the lecture with two hands-on examples in a language or in a software platform called R which is an open source free software for statistical data analysis. Now, before we proceed further, let me tell you that data analysis is an integral part of any research work. In every type of research that one carries out, invariably, one collects some data or the other. Even if **it's** theoretical research, at some point in time, there is **a** bit of experimental work involved to corroborate the theory that has been developed. By

and large, if you look at today, the scenario is about data analytics. A lot of data has been generated not just in engineering arena, but every where, and therefore, there is a need to understand how to analyze data in a systematic manner. There are many, many software packages that can carry out analysis of different types of data. And one should remember that analyzing data is not just about throwing data into the software packages, and then, reporting some of the fancy graphs that are generated and some nice results that you get on your computer screen - that is not data analysis; **it's** just a very tiny part of that exercise.

As much as we are analyzing data, there is a theory to it, and of course, we are not going go over the theory of data analysis, but the procedure that one needs to follow **to** get meaningful results. First of all, we need to ask - what is data analysis?

(Refer Slide Time: 02:36)


Data Analysis

What is data analysis?

Extracting **useful, relevant and meaningful** information from observations
in a **systematic** manner

For the purposes of

- ▶ Parameter estimation (inferring the unknowns)
- ▶ Model development and Prediction (forecasting)
- ▶ Feature extraction (identifying patterns) and classification
- ▶ Hypothesis testing (verification of postulates)
- ▶ Fault detection (process monitoring)

 NPTEL

Arun K. Tangirala, IIT Madras Introduction to Research 3

Analysis of data is about extracting useful, relevant, and meaningful information from observations. As I said, we collect data in different spheres and different forms, whether **it's** econometrics, whether **it's** engineering or medicine or any other field, we have different types of data collected and the purpose of collecting data is basically to know something about the process; but, of course, **that's** not the only purpose. As I have listed here, there are several purposes to data analysis. It could be **commonly**.. common

purposes, for example, parameter estimation where we are inferring certain unknowns, and when I say parameters here, it is not just model parameters; these parameters could also refer to statistical properties that we are commonly interested in, such as averages - that is mean, variability, variants or standard deviation, median and **and** so on. So, the term parameter here should be taken in a very generic manner. Essentially **it's** an unknown that we are trying to estimate from data. Obviously, **that's** one of the common exercises in data analysis.

Then, of course, we have the ubiquitous model development. Many of us collect data to develop models between two or more variables. And why do we develop models? Largely for predicting the variable of interest. Technical term for prediction is also called **is also** forecasting in econometrics and social sciences; **it's** a very common term and **that's** a huge field in itself. And then, we have feature extraction. So, if you take, for example, biomedical data, such as ECG or EEG data. In ECG data whether we have seen or not, but ECG is something that we all get to see whether **it's** a part of our research work or not. **It's** related to our human health, and if you see, if you recall ECG is a train, is a pulse, train of pulses which has regular peaks in it and we would like to extract those features. A doctor who looks at ECG essentially searches for those peaks, and also searches for anomalies in those peaks, and those anomalies could be indication of an unhealthy heart, for example. So, feature extraction is very common, not just of course in biomedical, but everywhere.

And classification. Once we extract features, we would like to classify image. In image analysis, this is a very **very** common exercise. We try to classify people with similar colors, hair colors, similar eye color, **and** facial skin color and so on.

And then, central to all of this is hypothesis testing. What is hypothesis testing? **It's** a very, very common exercise in data analysis, where the researcher postulates the truth being something. As an example, suppose I collect ambient temperature data. That is outside my room, I take a sensor and collect ambient temperature for about may be an hour or so. And I come back and postulate that the average temperature during that one hour is 25 degrees **s** Celsius. Now **that's** a truth that I postulate for the average temperature. And hypothesis testing is about testing that claim that I am making whether

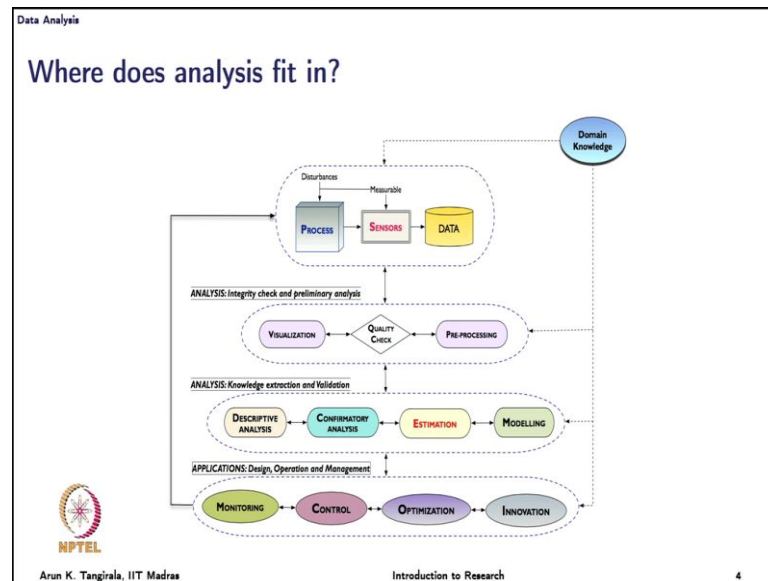
it holds or not, based on the data that I collected. So, the data I have collected serves as an evidence and I search for evidence in the data to support my hypothesis, and that has to be done in statistically sound manner. **And** hypothesis testing is an integral part of every statistical inferencing exercise. That involves, of course, a lot of theory, probability theory, estimation theory and so on. We **don't**, we will not go into depth in hypothesis testing and so on.

There is a MOOC course that we learn on introduction to hypothesis testing; that will begin perhaps in the second or third week of March. And if you are interested, you should sit through it because it is an integral part of every statistical data analysis exercise.

Then, we have data analysis being carried out probably for fault detection. As I explained earlier, a doctor, when the doctor looks at the ECG, is looking for anomalies or when a radiologist is looking at the x-ray, looking for anomalies based on some model or some template of normal conditions in mind, keeps projecting the present data on to the normal template. And that is how typically fault detection is carried out everywhere; even in process industry that is the situation. So, this is also known as Process Monitoring.

And there are several purposes that one can list in data analysis, but these primarily cover most of the applications of data analysis. As you can see the application are quite diverse and varied. Therefore, the procedure that we follow for an analysis, the tools that we deploy, and the interpretations that we draw have to definitely **vary** broadly.

(Refer Slide Time: 08:10)



So, **it's** useful to look at data analysis and ask where does it fit in, in the broad scheme of data driven analysis. So, this pictorial representation here tries to give you some idea of where analysis fits in. At the top of the schematic, you have process, and this process is a very generic process; it need not be an engineering process; it could be a social process or a biological process and so on. **And** we have sensors collecting data from the process and these sensors need not be just the classical hardware or instrumentation sensors. It could be also human sensors. When we collect survey data, we are the sensors, in fact. In the end, we have data with us and **that's** where the journey begins in data analysis.

In this module, of course, we are not going to talk about how to collect data, the techniques of sampling and so on; we are more worried about analyzing data. So, once I have data, typically I put it through certain steps and we will talk about this a bit later. The purpose of this schematic is to tell you where analysis fits in. The two stages that you see titled analysis here, it is essentially the representation of what goes on in data analysis.

For example, visualization is a part of analysis. When you plot something, and try to observe something from the data, that is also **a** part of the data analysis. **It's** a very important part or you may carry out parameter estimation as I just discussed or you may

build models and so on. Ultimately, the results from this data analysis go into some of these very broad applications as I just explained; it could be for process monitoring; **that** is fault detection or it could be for process control, where I am taking feedback action based on measurements. So, the controller is taking in the data, performing some kind of analysis, and then, deciding how the input should move. **That's also** a part of the applications of the data analysis or optimization, I am trying to come up with an optimized way of operating a process, or making innovations or changes to the process.

Then, of course, there are few other applications that I have not listed here, but ultimately you should understand that data analysis is not the end of the road, but it is really a very, very critical art of data driven process, analysis process operations and so on; and **when** I say process here, as I mentioned earlier, its fairly general. And in all of this, of course, the domain knowledge is very, very important as I highlight here, and we will talk about this a bit later when we return to the schematic and discuss the systematic procedure **that's** involved in data analysis. So, hopefully, you understand where data analysis fits in. It actually connects the world of observations to the world of information.

(Refer Slide Time: 11:15)

The slide is titled "Data Analysis" and "Types of analyses". It contains a list of five items: 1. Exploratory vs. Confirmatory, 2. Quantitative vs. Qualitative, 3. Descriptive vs. Inferential, 4. Predictive vs. Prescriptive, and 5. ... Below the list, it states "Exploratory and descriptive analyses are typically good starting points". A callout box at the bottom says "The objective of analysis determines the type of analysis and the tools that are used." The slide also features the NPTEL logo and the text "Arun K. Tangirala, IIT Madras" and "Introduction to Research" with a page number "5".

Data Analysis

Types of analyses

1. Exploratory vs. Confirmatory
2. Quantitative vs. Qualitative
3. Descriptive vs. Inferential
4. Predictive vs. Prescriptive
5. ...

Exploratory and descriptive analyses are typically good starting points

The objective of analysis determines the type of analysis and the tools that are used.

NPTEL

Arun K. Tangirala, IIT Madras

Introduction to Research

5

There are many different types of analysis, obviously, now, depending on the need and

the end use of your analysis, and there are several classifications that you can find in the literature. What I listed here is by no means exhaustive.

One of the common classifications of analysis that you see is exploratory versus confirmatory. In **an** exploratory type of analysis, you do not postulate anything upfront, but you are just exploring the data and trying to understand your process. For example, it could be looking at the mean or the standard deviation or looking at the graphs, trying to understand whether the process is linear or stationary and so on; trying to infer whether the process is oscillatory. There is a lot of visual plus some descriptive statistical analysis is involved; whereas in confirmatory analysis, you have a postulate upfront, and you would use the data to confirm whether what you have in mind is correct or not. So, obviously, there is a bit of contrast there.

And then, you have quantitative versus qualitative data analysis. In qualitative data analysis, for example, you look at trends; you are searching for certain features and oscillatory features, for example. Whereas, in quantitative you are being... you are looking at numerical results from your analysis.

Then, you have something called descriptive versus inferential. **And** in descriptive analysis, again, it is more or less like exploratory; whereas in inferential you are trying to infer something very specific from data and you are subjecting it to hypothesis testing. So, there is a lot of systematic, numerical, and statistical analysis involved in inferential analysis, and this could be a part of a larger analysis that you are doing such as building a model and so on. Whereas, descriptive analysis is not necessarily going to feed directly into a larger analysis; **it's** typically the first step in data analysis.

Then, you have predictive versus prescriptive. Predictive analysis is about asking where the process is heading based on data. And this is very useful in many, many applications; whereas prescriptive is about asking why something has happened. So, you have looked at the data either as a doctor or as an engineer or even as an econometrist **and** you looked at market data, and suddenly there has been a crash, and you are asking well – why **did** this crash occur or why did **a** particular fault occur in a process and so on. **And** then, you are, of course, going to suggest as to how to prevent this. So, there is going to be a

diagnosis and remedial thing involved in prescriptive analysis. So, obviously, there is a whole host of different types of analysis that are involved **and** there are many other classifications that I have not listed.

Typically, good starting points **s** are exploratory and descriptive analysis **and then** one proceeds from there; even if you have decided that you are going to build a model for prediction and so on, it is recommended that you explore the data, you make friends with the data, and you familiarize yourself with the data, before you proceed to developing a model, for example.

Now, in general, the objective of analysis determines the specific types of analysis that we have listed here, that would you carry out. That means, essentially what is the end use of an analysis, how are you going to use the information that you extract.

(Refer Slide Time: 14:44)

Data Analysis

Types of data

Several classifications (e.g., numerical vs. categorical, steady-state vs. dynamic, etc.) exist. In data analysis, the primary distinction is w.r.t the **data generating process** (DGP) or mechanism;

Deterministic (non-random) and **Stochastic** (non-deterministic) data

1. **Deterministic**: The DGP is perfectly known OR a mathematical function can perfectly explain the data OR there is no uncertainty in the observed values OR process is accurately predictable.
2. **Stochastic**: Observations are only a subset of many (countably finite or infinite) possibilities OR no mathematical function can explain the observed data OR observations have uncertainties in them. Stochasticity does NOT necessarily mean zero predictability

In practice, no process is deterministic since perfect knowledge is never or rarely available.

NPTEL

Arun K. Tangirala, IIT Madras

Introduction to Research

6

And before we proceed to data analysis, it may be very useful to know what are the different types of data that are available, because primarily as I will mention later, the tools, the interpretations, and so many other steps depend on the nature of the data. Now, once again, when you turn to data analysis literature, you will find different types of classification, a number of classifications **s** exist. For example, you may find numerical

verses categorical classification or steady state verses dynamic and so on, but I am not going to really talk about those, because **that's** definitely important, but what is more important which people often ignore, at least beginners, in data analysis often ignore is this classification of deterministic verses random or stochastic data. **That's** very, very important because theoretically itself or **the** theory itself largely differs the moment you say the data has come out of a deterministic verses stochastic process.

Now, of course, some people use the term random. There is **a subtle** difference between random and stochastic; however, we will not observe that distinction here. Sometimes, the term non-deterministic is used. Before we ask why this classification or distinction is important, **let's** look at the definitions and the definitions are based on the data generating process. So, the deterministic and stochastic qualifiers are not for the data per se, it is actually, they are qualifiers for the process that is generating data. In data analysis, the data generating process is **a** very important entity. What we mean by data generating process is the process that you are collecting data from as well as your experimental process. That is how you are conducted the experiment. So, put together, everything is the data generating process. **It's** not just the physical process alone that you are analyzing. **That's** very important to remember because both contribute to the type of data that you generate.

As an example, let us say, I am observing some reactor temperature which is being controlled by **coolant** flow. This is very common in nuclear reactors and so on. Now, the reactor that I am observing is a physical process. There is a controller there trying to control the temperature and so on. **Let's** say **it's** doing a fantastic job of controlling the temperature. So, the temperature is held at its set point, but the sensor that I am using can be quite noisy. Ultimately, the data that comes to me is not directly from the process between the process, the physical process and the data I have a sensor and that sensor characteristics. The characteristics of that sensor actually goes into the data. If the sensor is very noisy, you are going to end up with **a** noisy data. If the sensor is **a** biased, your data will show a bias.

Now, the data generating process consists of both the reactor and the sensor. The DGP - as we shall use as an abbreviation that you will find even in literature - is set to be

deterministic, if you can find a mathematical function that can accurately explain the data; that is, it can predict the data, you can fit in perfectly, every data point lies on the curve of this mathematical function. That's one way of looking at it or you can say that you have exactly understood the data generating process; there is nothing that you cannot explain or that you can say that there is no uncertainty in the observed values.

You may have... many a times what happens is you may have very good knowledge of the physical process. For example, the reactor in the example that we discussed earlier, I may understand the dynamics, the kinetics, everything in the reactor, the thermodynamics and so on, but I may have poor knowledge of the sensor characteristics. Now, that makes it a very tricky situation. It puts in neither in the deterministic nor the stochastic thing and we will talk about it, but essentially that is not deterministic anymore. The data is not coming out of a deterministic process because a part of the DGP, which is a sensor, has not been understood properly. In general, this is the situation; sensors are never perfectly understood. You can design a very, very nice sensor that has very less noise, but no sensor is perfect and no sensor is really perfectly understood. So, in reality, there is nothing like a deterministic process. It's an idealization, that is very useful in data analysis and in developing the theory.

Now, obviously, stochastic is the counter part for the deterministic, where we say that the DGP is not accurately predictable or its not known or you can say there is uncertainty in the data and so on. And as I said, in practice, no process is deterministic since perfect knowledge is not known. Now, does it mean that every data generating process has to be treated as stochastic? Not really. There is a certain distinction still observed between deterministic and stochastic processes, even in practice even though no process is perfectly predictable. Typically, we say that a process is predominantly deterministic or predominantly stochastic and we will talk about it very soon; soon after this slide here.


(Refer Slide Time: 20:26)

Data Analysis

Why is this assumption important?

The **theory**, tools for analysis and **interpretations of the results** heavily depend on the assumptions of determinism or otherwise.

- ▶ Example: definitions and computation of periodicities, mean, variance, stationarity



Arun K. Tangirala, IIT Madras

Introduction to Research

7

So, going back to the question that we raised earlier - why should I worry about **whether** data has come from a deterministic or **a** stochastic process? Why is this assumption so important? Now, the answer is something that I have said earlier - that theory, the tool for analysis, and the interpretations of the results significantly depend on these assumptions. And there are many, many examples that one can give. For example, if I am looking at periodicity detection, then first I have to define what **is** a periodic deterministic signal is. If I am looking at deterministic data, and if I am looking at stochastic data, then I have to define what is meant by periodic random process. So, the definitions vary, and therefore **right**, the interpretations will also change. Likewise, when I am referring to the average of a stochastic signal, normally misconception is the sample mean, which is the mean that is computed from observations, is considered as the mean of the process itself; that is, it is not true at all. The sample mean **is** only an estimate of the true mean which is defined in a different way and we will quickly talk about it.

So, the bottom line here is, whether you **are** looking at mean variants or any other characteristic of the signal, you have to really go back to the books and ask how are these defined for the signal that I am looking at. If I am looking at a deterministic signal - what is the definition of periodicity? **What's** the definition of mean? If I am looking at **a** stochastic signal - what is **the** definition? And then only, you can make meaningful

interpretations. **And** a very, very common situation is that of power spectral density which is used for periodicity detection and so on. Anyway, we will not go so much in detail, but you should remember this fact, and every time you collect data and you sit **do** analyse data to ask certain important questions, and one of those questions is - what assumptions am I making about the data generating process.

(Refer Slide Time: 22:35)

Data Analysis


In practice: deterministic or stochastic?

Q: When do we assume the DGP to be deterministic or stochastic? Can we know a priori?

- ▶ Determinism or stochasticity (random) rests with the knowledge of the user (analyst)
- ▶ Whenever the causes are unknown or the underlying mechanism is too complex to be understood, it is reasonable to assume the data to be stochastic (e.g. sensor noise, econometric data).
- ▶ Where the physical process is well-understood and/or the causal variables are known, and the errors in observations are "negligible", data may be assumed to be (predominantly) deterministic.

Q: Is it possible to have a mix of both situations?

- ▶ Yes. Several processes fall into this category (e.g. engineering data).

 NPTEL

Arun K. Tangirala, IIT Madras

Introduction to Research

8

So, before we move on to learning what is a procedure **that's** typically followed in data analysis, **it's** again good to close this discussion on deterministic or stochastic with this question - when do we assume **the DGP** to be deterministic or stochastic? Does some one comes and tell me that? Obviously not. Can I find in the literature? Some times yes, but most of the times no for the process that I am looking at. So, is **there** a way to know a priori? Well, unfortunately, there is no perfect formula for you to figure out whether the data is coming out of a deterministic or stochastic process; **but** you have to go by how much you know vis-a-vis how much you do not know **right**. **That's** about the process.

So, as an example, suppose I hold a ball, in a room, where there is not much breeze. The air is fairly still and I am going to drop the ball. We can pretty much with a help of physics calculate how long the ball takes to hit the floor. **That's** fairly a deterministic process. There is not so much uncertainty there. On the other hand, if I were to be asked

whether its going to may be rain tomorrow or in 6 months it's going to rain or if I look at any other atmosphere process characteristics or even econometric process and so on or a sensor noise and so on, **it's** very hard to predict **what is** the next value of the error in the sensor, for example.

In such situations, we take asylum in probability theory and we say that look I do not have a perfect knowledge of the process. Therefore, I shall assume that there are many possibilities, and I assign chances to each of those possibilities, and **that's** where probability theory takes birth, and **that's** where time series analysis comes in. There is also another situation in which you may have to assume the process to be stochastic and that situation is where the cause of the data that you **are** looking at is unknown. So, as a simple example, suppose, I am looking at the prediction of stock market index. There are so many causes that actually affect; there are so many factors responsible for stock market index. Now, unfortunately, even if I know the causes, I may not be able to measure them and sometimes I **don't** even know the factors exhaustively enough.

So, what do I do? I depend on the history, and hope that history has some reputation in it, and I exploit the historical patterns and correlations and so on, and make a prediction, but that prediction is going to be, obviously, not accurate. The point is here we apart from looking at the history, the main assumption that we make is that this stock market process – **that** which is some ficticious process that generating the index - is random in nature. It throws out, it can throw out any value of the stock market index out of which I am observing one. There could have been many other possibilities out of which one has occurred. In reality that may not be the case. The true process may not **be behaving** that way; it is an assumption that I am making, so as to move forward in data analysis; **that's** very, very important. In other words, whenever I assume a process to be stochastic, it is a formal way of saying I do not have sufficient understanding of the process.

But there are situations where I have **a** mixed case - where I have a fairly good knowledge of a part of the process and the other part of the process is beyond me. That's where we have a mix of deterministic and stochastic process. And that is very common in engineering arena because a physical process may be well understood, like in the reactor example that we discussed earlier, but the sensor or some effects of unmeasured

disturbances, something **that's** beyond my control, those have to be treated as stochastic simply because I can not predict them accurately.


So, there we may have to deal with both the deterministic and stochastic comments, and there is a theory that allows you to do that. So, the bottom line is please do spent a few moments asking - first of all whether you want to treat the data as a deterministic, coming out of a deterministic or stochastic process or a mixed process, and you should have the answer as to why you have made that assumption, because the course of analysis really depends upon that.

(Refer Slide Time: 27:18)

Data Analysis

Further

- ▶ Be clear on why the data is assumed to be non-deterministic.
- ▶ Suppress the sources of randomness, if possible, during experimentation.
- ▶ **IMPORTANT: Any estimate inferred from data with uncertainties also has uncertainty in it!**
Therefore, always attempt to **compute the size of the error in the estimate.**
- ▶ Choose an estimation method that delivers estimates with the lowest possible error and improves with the sample size N , especially error goes to zero as $N \rightarrow \infty$.

 NPTEL

Arun K. Tangirala, IIT Madras

Introduction to Research

9

Okay so, just to emphasise, be clear **on** why the data is assumed to be non-deterministic, particularly this is a very common assumption; in fact, many **many** researchers carry out data analysis assuming the data to be stochastic without probably **even** pausing for a moment to ask why and what kind of stochastic. Just because I assumed data to be coming out of a stochastic process, I cannot simply go ahead and carry out my data analysis. Within the stochastic processes, there are several classifications and I have to be clear in my assumption. It is after all an assumption, but still I have to be clear as to whether **it's** coming out of a stationery stochastic process, non-stationery stochastic process or a periodic stochastic process and so on. So, there are different categories and

we should be very clear. **This is**.. all of this is apart from the regular classifications that we have to think about - which is whether the data is coming out of a steady state or dynamic process, linear or non-linear. All those still hold good, but this is a very important classification that you have to make upfront. And, in general, if you have access to the experiment, this is, of course, outside the purview of data analysis, but **it's** related to suppress the sources of randomness if possible during your experiment. That is a part, that is a topic that is dealt within detail in design of experiments and we **don't** go into that here.

The most important thing that people often miss out on - that is beginners - is that any estimate or any inference that I draw from data - and typically lot of data is assumed to come out of random or a stochastic process; those estimates are inferences are also random in nature, because you are putting through data which has uncertainty in it. Through some mechanism, some formula, some mathematical process and you are getting out, for example, I am computing the sample mean. I go and collect data, ambient temperature data, with a sensor.

Come back to my desk and feed in the data into my computer or my calculator and I actually compute the sample mean. The sample mean that I have calculated is simply the average; average of what? Data that is corrupted with uncertainties. So, those uncertainties have also made their way through to your sample mean. So, the value of sample mean that you have has also some uncertainties. The uncertainty here is not with respect to your calculation error **okay**; that should be kept aside. The uncertainty in it has got to do with a fact that you are not sure if this is **the** true mean of the process; that is the uncertainty. **That's** again because the data that you collected **is has** ..is one of the many possible data sets that you could have collected. So, you have a sample mean which is one of the many possible sample means that you could have obtained. And this is true for any parameter estimate whether you are estimating slope, intercept or any other correlation, any other complicated parameter that you may be estimating from stochastic data, you will have to remember that there is an error in the data, and as an analyst, **it's** the responsibility of the analyst to compute the size of the error or the uncertainty in the estimate or the inference that you are drawing. **That's a** very important part of data analysis. And in this context, you should choose an estimation method that

gives you an estimate with as small error as possible and we will talk about it a bit more in detail.


More importantly that as you collect more and more observations, the error in your estimate should ideally start shrinking and go to zero asymptotically as n goes to infinity. So, these are some criteria and very often many of the researchers may not be aware of this.

(Refer Slide Time: 31:29)

Data Analysis

Terminology in data analysis / statistical inferencing

- 1. Population and Sample:**
Population refers to the collection of all events or possibilities, typically also known as the ensemble. Sample refers to the subset of population (observations) that has been obtained through sampling.
- 2. Truth and Estimate:**
True value refers to the population characteristics in descriptive statistics or to that of true parameters / signals in parameter / signal estimation. Estimate is the inferred value of the unknown parameter / signal from observations.
- 3. Statistic and Estimator:** A statistic is a mathematical function of the given observations. An estimator is also a mathematical function of the observations, but devised for the purpose of inferring or estimating an unknown quantity / variable.

Arun K. Tangirala, IIT MadrasIntroduction to Research10

So, before we move on to the data analysis procedure which is the last part of this lecture, it's important to at least gain familiarity with some terminology that's commonly found in the data analysis literature. And here, we will assume primarily that we are going to deal with data that comes from a non-deterministic process, because that's the prevalent situation everywhere. Even if you have a mix of deterministic and stochastic process, these terminologies apply.

So, the first terminology that we have to be familiar with is the term - population and sample. We know that in stochastic process, as we just mentioned earlier, we observe when we get a data record, the record of data that we have is one of the many, many possibilities - sometime countably finite, sometimes countably infinite or infinite - many

possible data records exist, out of which we have observed one; that is the theory, the entire theory is built on that premise. The population refers to all possibilities that you could have seen, although we do not observe that; it is not possible to really have or observe all the possibilities in most of the situations. So, population refers to the collection of all events or possibilities.

When with respect to the temperature measurement that we talked about earlier of the ambience. If I collect hundred observations with one sensor over a period of time, another sensor would have viewed that in a different manner; so that, another data record could have been impossible or the same sensor itself could have recorded, everything held constant **could** have recorded completely different values. So, **it's** not possible obviously to realize all those possible values, but theoretically it is important to imagine this all possible space and we call it as population of which we obtain one sample. When I say sample here **it** is not one observation, a sample in statistics typically refers to collection of observations. It basically refers to one data record, but in signal processing, a sample could be one value at an instant and that is in the deterministic world. Corresponding to this population and sample we have a truth and an estimate. We just said population is a collection of all possibilities, but **that's** characterized by the truth.

So, this entire population has, for example, certain mean. Let us say, I am looking at the average salary of a group - of an age group - in a very large country, let us say or in very large firm. **It's** not possible for me to go and ask every one. For practical reasons, I may only select a few people to ask to collect this information about their salary and from which I infer the true average salary. So, the true average salary is a truth, that is what characterizes the population - that age group - and estimate is what I obtain from data. Obviously, estimate will always be different from truth because I have not seen all possibilities. And the estimation theory tells you how to estimate the parameter of interest, in an efficient consistent way, from the data recorded. Associated with this, we have this estimation. You have this terms called statistical **and estimator**. A statistic is nothing but a mathematical function of your observations. Again, the sample mean is a statistic; sample variance is a statistic. **It's** simply a mathematical operation that you are performing on the observations. An estimator is also a mathematical function. You can say **it's** a soft device or virtual device that operates on the data in a mathematical way

and gets you an estimate of what you want. The difference between statistic and estimator is fairly **suttle**. Prima facie they appear to be the same, but the difference is, statistic is merely a mathematical function not necessarily devised for any purpose; whereas, an estimator is a mathematical operation that is carried out with a specific purpose of estimating something in mind **okay**.


So, if you go down further in the estimation theory, you will see more discussion on this difference between statistic and estimator, but as far as the mathematical operation, both are the mathematical functions of the data. So, in estimation theory, we use the statistics and estimators quite often. So, sample mean is both a statistic and an estimator of the true mean **alright**, because sample mean is devised for estimating the true mean, and again associated; as you can see **it's** all about the population and sample; you have ensemble and sample averages.

(Refer Slide Time: 36:47)

Data Analysis

Terminology

- 4. Ensemble and sample averages:**
Ensemble average is the average computed across all set of possibilities (population), at a single point in time (or another domain). Technically this is known as **Expectation**. Sample average is the average that is computed over all observations for one data record.
- 5. Bias:** It is the systematic error in an estimate $\hat{\theta}$ of a parameter θ introduced by virtue of the estimation method. Bias is a measure of **accuracy**. Mathematically, $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta_0$, where $E(\hat{\theta})$ is the ensemble average of estimates across all data records.
- 6. Variability:** It is a measure of the spread of estimates obtained from all experimental records. Variability is a measure of the **precision**. Mathematically, $\text{var}(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2$, i.e., the ensemble average of the squared errors in the estimate across all records.

 Ideally one would like to obtain estimates with zero bias and as low variability as possible!

Arun K. Tangirala, IIT Madras Introduction to Research 11

Ensemble average is the average that you would compute from all possible values that you could have ever seen, which, of course, we never get to see; in most of the situations we never get to collect all the possible data records or in the salary example, it may not be possible to collect the data from all possible people in meaningful time, Sample average, of course, is simply the average is computed from these samples. And these

averages are used to define what is known as **an** ergodic process and so on; we will not go into that.

Finally, the two terms bias and variability are very important in data analysis because they both talk about the error that one makes in estimating some parameter or a set of parameters. Bias is a systematic error that one ends up making; every estimation exercise will give you an estimate **that's** in error; **there's** no doubt about it. The question is whether the difference between estimate and truth is by chance or due to some deficiency or shortcoming of your estimator.

How do you know if there is bias in your estimation exercise or in your estimate? Imagine this thought process; I have collected let's say hundred data records or thousand data records or thousand data records for one process, and I compute the sample mean for each of the data records. If I average, let us say, not even thousand, **about** a million data records, per data record I have one sample mean; and for million data records, therefore I have million sample means. If I average the sample means, the million sample means that I have, I should be very, very, very close to the truth. If there are only million possibilities, I should get the truth, but if there is a difference that still exists between the average of this million sample averages or sample means and the truth, then there is a systematic error. This ensemble average that we talked about, as I have written above, is also known as expectation. **So** in other words, if coming back to bias, if the expected value of the estimate is different from the truth, then we say that there is a bias in the estimate.

Now, coming to variability, **it's** a measure of the spread of estimates across different data records. **It's** very, very important because in practice I only work with one data record and I make all my inferences from a single data record. I need to know, I need to know the answer to the question - had I collected another data record, and had I drawn the inferences from the data record, would those inferences be significantly different from what I have seen now. They will be different, obviously, numerically, but will the difference be too large? And if that is too large, then we say there is a large variability in the estimation or in the estimate. For this reason, variability is also referred to as precision. The lower the variability in the estimate, the more precise is the estimator. In

contrast, bias measures accuracy. Lower the bias, lesser the inaccuracy. In fact, zero bias would mean, it's an accurate; the estimator is an accurate one gives you an accurate estimate of the truth. Typically, we are more worried about variability. Yes, of course we want an accurate estimate; ideally, we want an accurate and a precise estimate. It's not possible to obtain both an accurate and precise estimate from finite samples, but what definitely we would like to have is low bias or zero bias if possible, and as small or low error as possible. That is why you see on many instruments that are being supplied to laboratories, you have these instruments being qualified as high precision instruments.

Remember, estimation is not just about we using a mathematical formula on data. Even a sensor actually, the reading that it gives you is an estimate of the truth. So, one should have a very broad view. So, what the company or the manufacturer means by high procession instrument is, if you were to repeatedly measure the same variable with the sensor, you would not see too much variability. Definitely there will be some variability. So, take a thermometer, and place it on your body, and take the reading. Note down the reading. Then again, take the thermometer, then place it on the same part of the body and take the reading. Of course, you will probably see the same reading there, because thermometers don't report beyond first decimal, but you may see an actual variability if you have thermometers reporting may be the fourth or fifth decimal and so on. That is called resolution of the instrument. But, anyway, the point is repeated measurement of the same variable definitely will give you different readings, but you don't want wide variations in those readings. So, you want a high precision instrument.

So, with that we come to the close terminology or the major terminology. Of course, there are many other terms that one would want to be familiar with, but those are the basics.

We will conclude the lecture with the procedure that one needs to follow for data analysis, and then, move on to a couple of illustrative examples in R.


(Refer Slide Time: 42:26)

Data Analysis

Data analysis procedure

Almost all data analysis exercises are **iterative**, not necessarily sequential.

The success of any data analysis exercise depends on two factors,
quality of data and **methodology of data analysis**.



Arun K. Tangirala, IIT Madras

Introduction to Research

12

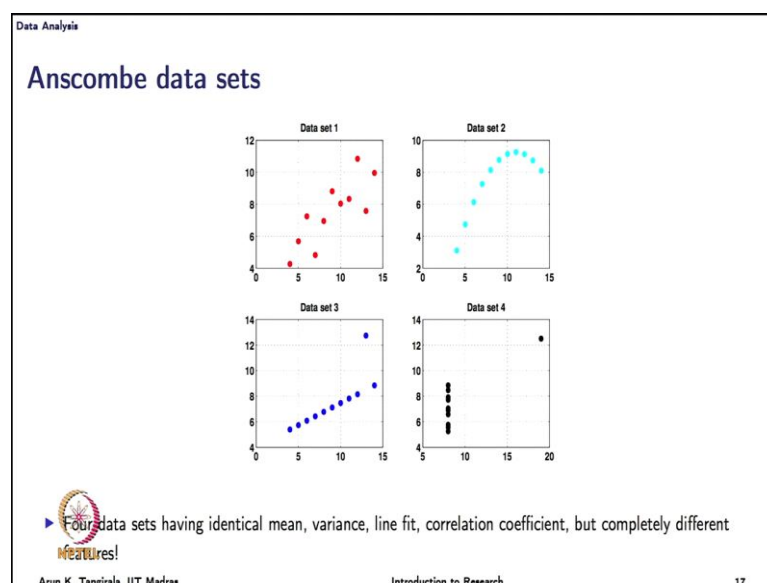
The foremost thing to remember is that almost all data analysis procedures are iterative in nature. **Don't** expect step 1, step 2, step 3, step 4, done. **It's** never going to be that way, very rarely. So, they are not necessarily sequential, but there is a certain order in it. It **doesn't** mean that you go about doing in a haphazard manner. And also, remember that the success of any data analysis exercise depends on two things: the quality of data that you collect and the systematic procedure or the methods that you have, the tools that you have used in data analysis or the entire method methodology of data analysis that you followed **ed**. If the data quality is bad, you cannot do much. Even if you follow systematic procedure, you may not get marvellous results; and even if the quality of data is good, you follow a very unsystematic procedure and you use a very poor estimator, you may end up with a poor estimator. So, both are important.

Now, of course, quality of data is many a times is not in our hands, but you should still be aware of the theory to be able to say when you look at a data and say that well this data quality is not good, so I should not expect good results. Suppose, you are the consultant, and you have been given some data, you are hired for the data analysis by a process industry and you are given some data, the first thing that you should be doing **is** look at a quality of a data and make an intuitive assessment of how good are the nature of the results that you can expect. If the data quality is too bad, you should return the

data and say well **don't** expect miracles. The data has to be collected in a much better manner, so that is so-called informative **alright**. So, please keep that in mind.

Now, we revisit the schematic that we saw at the beginning of this lecture which gives us some idea of what is the systematic procedure of the data analysis. So, at the top, we have, of course, the data coming out of the process - well, out of the sensors - and the first step in data analysis is going through some preliminary steps which involves **three** things: visualization, quality check, and pre-processing. Now, again within this, **it's** not necessarily sequential. Generally, the first step is to plot our data because the numbers may not tell you much, but pictures tell you **a** lot more and to emphasize that let me actually take you to the slide here.

(Refer Slide Time: 45:06)



This is a very classical example based on what is **known** an Anscombe data set. And this data set is also available in **R** as many data sets that come with it when you install. Now, the beauty of this Anscombe data set is... let me explain what I show you here. I have four data sets here, of pairs of variables; think of $x_1 y_1$, $x_2 y_2$, $x_3 y_3$ and $x_4 y_4$. They all look visually very different; however, if you had not looked at the data, and you **had** computed, for example, mean or variance or you had fit a model between the respective pairs or compute at a correlation, they would all be identical, they would **all** be

identical. So, based on the descriptive statistics or even your regression, you would not be able to make any distinction between these four data sets. However, as you can see, there is such a wide difference in the features of each of this data. For example, if you look at the data set 2, there is a parabolic kind of relationship, whereas data set 4, all the points fall on a straight line. Practically, there is no change in x, only y is changing apart from one outlier. So, this data set is also very useful to show you what outliers can do to your data analysis.

Outliers are those data points that fall out of the normal thing or that fall out of the region where the rest of the data. In fact, if you look at data set 3, all the data points lie on a straight line except for one data point **that's** an outlier **right**. So, two data sets can look completely identical when you look at descriptive statistics alone, whereas the moment you plot them, you know there is a huge difference. And then, of course, we will take corrective action and the necessary steps to be able to extract the features correctly. So, that hopefully **this** drives home the need for visualizing data prior to any data analysis.

And then, of course, we just did a quality check. For example, we observed the presence of the outliers. **That's** very important. Many a times data come with missing values. You have to decide how to deal with missing values. There is a whole lot of theory available whether you should replace it or live with it, interpolate, and so on. **So** obviously, it's not possible to talk about it here.

And then, other kinds of pre-processing that you may have to do. For example, you may have to sync the time stamps in multivariable processes; where you collect data from different variables, you may have to sync the time stamps or sometimes one variable is collected at one sampling rate, another variable is **collected at** an another sampling rate, then you may have to determine how to pre-process.

Sometimes, in fact, lot of times data comes with noise and we may have to remove the noise prior to any other step that you we may take **as** analysis, because noise really can hijack the inferences that we draw. It can significantly affect the course of data analysis. So, **we** may want to filter out noise.

So, there are many, many reasons why you want to spend time in this preliminary step which is perhaps **the** most important. Some bit of ... there is a bit of drudgery involved in it. It can really get on to your nerves, and there is a golden **80/20** rule which says of the 100 percent of the time that it takes for overall data analysis, you may have to spend 80 percent of the time on data pre-processing. **That's** the kind of effort that you may have to put in sometimes. If you are lucky, of course, you can just go pass through this step in a breeze.

Then comes, of course, the core part of your analysis which is modeling or estimation or exploratory or may be predictive kind or prescriptive kind of analysis diagnosis and so on, the results of which feed into many applications, such as monitoring, control, optimization, innovations as we have discussed earlier.

Now, very importantly, in all of these steps, you should keep in mind two things. One as the arrows indicate, they are not necessarily iterative. That means, sometimes you may have to go back. So, if I find that the data quality is poor, I may have to go back to my experiment or the person who gave me the data to give me better data or that I find that the features that I want are not present in the data, I may have to redo the experiment or after estimation I figure out that the data quality is not good enough or that I **had** not done sufficient pre-processing, I may have to go back and so on. That is the first point.

And the second point is if you have any domain knowledge, and **that's** very, very important; obviously, you should know where the data is coming from, whether it is coming from a biological process, an econometric process or an engineering process or some social media process **and** so on... or a chemical process and so on. That knowledge should be factored in at every step even in choosing what should be the sampling rate for your data - that is how often you should collect the data; **that's** where your journey of domain knowledge begins.

And if I know, for example, I am looking at a PH system, I know **it's** a non-linear system **right**. Then when at the modeling stage, I should not even think of building linear models, but if I **don't** know where the data is coming from, it just remains a data set and its of limited use to me, its just some kind of a toy that I am playing around with. So, the

domain knowledge has to be used at every stage in the best possible way and that completely depends, of course, on how much domain knowledge you have. It may help you in choosing the structure of the model, the kind of features that you should search for or the level of noise that you can expect - whether it is a sensor noise or a process noise and so on, or even in your applications. So, that should be kept in mind. All-in-all iterative in nature, but there is a certain **systematizm** to it - choose your models properly and make sure that you go through the data quality check; visualize the data; pre-process; then choose your models properly, if you are building models and so on. We will talk about that in the next lecture on modeling skills or estimate properly and so on.

So, we will now expand on this a bit more. So, before we proceed to data analysis, the first step, as we learned, is to ask certain preliminary questions.

(Refer Slide Time: 52:08)

The slide is titled "Data Analysis" and "Preliminary questions in analysis". It contains a list of five questions with their respective answers:

- ▶ **What type of analysis is required?** Objectives or purpose of data analysis (e.g., exploratory or confirmatory, descriptive or predictive)
- ▶ **Where does the data come from?** Sources of data (domain)
- ▶ **How was the data acquired?** Method of data acquisition / collection (sampling mechanism, hardware used, etc.)
- ▶ **How informative is the data?** Quality of data (e.g., outliers, missing data, anomalies, noise levels)
- ▶ Lastly, but perhaps the most important, **what assumptions are being made on the data generating process?** (e.g., deterministic / stochastic, linear / non-linear, stationary / non-stationary)

The slide also features the NPTEL logo and the text "Arun K. Tangirala, IIT Madras" and "Introduction to Research" at the bottom.

What type of analysis is required? Whether it is descriptive, predictive and so on, because that changes entire course of analysis. Where does the data come from? Sources of data - we have just asked, we have time domain, data frequency domain, data biological process data and so on. How was the data acquired? What kind of hardware was used? Was it a human who collected the data? Was it a sensor? If it was a sensor, what was the sampling rate? So many questions may have to be known for us to be able

to say how good a quality of analysis that I can expect, and what should be the kind of analysis tools that I should be using, and what should I be searching for. And then, of course, how informative is the data – **that's** a bit qualitative, but there are also certain quantitative measures of information. We will not talk about it, but as I said, it basically pertains to quality check whether there are missing data outliers and so on.

And lastly, but perhaps the most important - what assumptions are being made on the data generating process should be clear. You may be wrong. In the beginning, we may make wrong assumptions, we may assume a stochastic process to be deterministic, we may assume a non-linear process to be linear, a steady state process to be dynamic and so on; **it's** possible. But we should stick to the assumptions that we have made throughout the data analysis, and choose and be the consistent and commensurate - that is we should choose right tools and so on, so that if you have made a mistake, in the end, we do know that we have made a mistake. And then, we go back, and make corrections. **And** I am not going to really go into this detail. We have already talked about it. So, once you are through the preliminary questions, we prepare our data, through many data pre-processing steps; depends on what kind of problems the data has.

(Refer Slide Time: 54:00)

The slide is titled "Data Analysis" and "Preparing your data". It contains three main sections:

- Data pre-processing**
 - Cleaning up the outliers or faulty data, re-scaling, filling in missing data
 - Robust methods can handle outliers or faulty data without an explicit treatment
 - Data reconciliation - adjusting data to satisfy fundamental conservation laws
 - Synchronization of time-stamps in multivariate data
- Partitioning the data**
 - Partition into "training" and "test" data sets.
 - In certain applications, several data sets may have to be merged.
- Assessing quantization or compression errors** for data from historians.

The slide also features the NPTEL logo and the text "Arun K. Tangirala, IIT Madras" at the bottom left, "Introduction to Research" at the bottom center, and the number "15" at the bottom right.

For example, if the data is outliers, I mean, I have to clean up, but there are also robust

methods. We say let the outliers stay in the data and still going to get you decent estimates and so on, or **we** may have to filter and so on.

The one thing that I **didn't** mention earlier is partitioning the data. If you are building a model or if you are performing classification by extracting features and so on, always you should have a part of the data in hands for cross validation. We should take the model that you trained on the data and test it on a fresh data which has come out of similar experimental conditions, or if we have built a classifier based on the training data, then we should test the classifier on the part of the data set that I have updated. So, always keep portion of the data set for cross validation.

And sometimes it may be necessary to assess quantization or compression errors that may arise in data that come from historians. These days it may not be so bad, but if you are looking at data may be 10 year ago, and so on, they be **in** highly quantized or compressed, and that can significantly change the course of your data analysis. They may introduce unnecessary non-linearities which **are** not a part of the physical process. **It's** just a part of your data acquisition process.

(Refer Slide Time: 55:34)


The slide is titled "Data Analysis" and "Visualization". It contains a quote: "Always convert numbers to pictorial representations before, during and after mathematical operations - visual analysis reveals a wealth of information". Below the quote is a list of visualization techniques: Simple trend to high-dimensional plots, Bar charts, Box plots, histograms, Color maps, contour plots, Specialized plots, and The slide also features the NPTEL logo and the text "Arun K. Tangirala, IIT Madras" and "Introduction to Research" with the number "16".

Data Analysis

Visualization

Always convert numbers to pictorial representations before, during and after mathematical operations - visual analysis reveals a wealth of information

- ▶ Simple trend to high-dimensional plots
- ▶ Bar charts, Box plots, histograms
- ▶ Color maps, contour plots
- ▶ Specialized plots
- ▶ ...

 NPTEL

Arun K. Tangirala, IIT Madras

Introduction to Research

16

And of course, we talked about visualization where in we could look at simple trend to

high dimensional plots, bar charts and so on. The list is obviously not exhaustive, but always convert numbers to pictorial representations before, during, and after analysis. So visualization is a part - integral part - of every stage of the data analysis. So, we have talked about the Enscombe data sets that basically emphasise a need for visualization.

(Refer Slide Time: 55:48)

Data Analysis

Core steps

- ▶ **Select the domain of analysis** (e.g., time, frequency).
 - ▶ Transformation of data does wonders (e.g., Fourier transform for detecting periodicities)
- ▶ **Dimensionality reduction** for large-dimensional data
 - ▶ Multivariate data analysis tools (e.g., PCA, NMF)
- ▶ **Choose the right mathematical / statistical tool** appropriate for the purpose of analysis and commensurate with the assumptions made on the data.
 - ▶ In hypothesis tests, the appropriate statistic and significance level should be chosen
 - ▶ In analysis of variance (ANOVA) problems, sources of variability should be clearly identified.
 - ▶ In modelling, appropriate regressor set and model structure must be chosen.
- ▶ **Always factor in prior and/or domain knowledge** at each step.

NPTEL

Arun K. Tangirala, IIT Madras

Introduction to Research

18

The core steps in analysis is what we will talk about and then conclude the lecture. Select the domain of analysis. What we mean by domain here is whether we want to analyze the data in the same domain that it has been obtained. Normally the time is a domain - independent domain - in which we collect data, but we may not want to analyze the data in the time domain. Sometimes we may want to move to frequency domain. In fact, that's a very common procedure that's followed in periodicity detector. If you look at atmospheric data or even data coming from vibration machinery, ECG or EEG data, for all such data, the frequency domain analysis is a very convenient domain of analysis because it reveals the features that can be directly related to physical characteristics of the process. And of course, many a times you have to do joint time frequency analysis. Wavelets is one of them. So, select the domain of analysis and time need not be, it could be on space also. Many a times data can come to you as a function of frequency straight away, like observance spectroscopy data comes as a function of frequency from the chromatograph.

Then the second one that we may have to take a decision on is dimensionality reduction and particularly when we are looking at large dimensional data. This is now increasingly becoming common. It's not possible to analyse multivariable data in large dimensions, particularly when it goes to hundreds and thousands and so on, and they are very common today. And machine learning algorithms come very handy in this respect. Principal component analysis, non-negative matrix factorization, and many other dimensional reduction techniques are very powerful in analyzing large dimensionality data, large dimensional data. So, we have to take a decision whether to analyze in the dimensions that in which I have obtained or in a reduced order or a lower order dimension.

And most importantly is to choose a right mathematical or statistical tool. For example, if I want to estimate the mean, I can choose mean, sample mean or sample median as an estimator. Now, when should I choose sample mean as an estimator, when should I choose sample median - there is no perfect answer for it, but there are some very good guidelines. If I know that the data is free of outliers and it's coming out of a Gaussian distributed process, sample mean gives you the most efficient estimator - meaning estimate to the lowest possible error among all the estimators; but if I note that the data could be corrupted with outliers and I don't want my estimator to be affected by it, then sample mean is not robust. That means it's very sensitive outliers, then I should use the sample median. This depends on prior knowledge that I have and the knowledge that I have acquired from my preliminary data analysis step. And these kinds of decisions have to be made at every stage and there may not be unique answer.

Remember that; in data analysis, there is no unique answer necessarily at every stage of your analysis and even towards the end, even when you build a model, there may be many models that can explain the given data and you may have to actually pick the model based on some criteria. And likewise, in hypothesis test, we have to choose an appropriate statistic and analysis of variance. We may have to first know what the sources of variability are; at least guess fairly well enough or in modeling as I said you may have to choose the right regressor set and model structure and so on, and always factor in prior and/or domain knowledge. You may know a priori something, not necessarily related to the physics or the process, but may be from the sensor

characteristics and so on, and of course, you may know something about the domain or the physics or the chemistry or the biology of the process that you can easily factor in and save a lot of headaches.

(Refer Slide Time: 60:11)


The slide is titled "Assessing and reporting your results" and is part of a "Data Analysis" presentation. It contains a list of five bullet points: 1. Compute confidence regions and errors in estimates. 2. Conduct hypothesis tests on estimates. 3. Cross-validate models on fresh data sets (did the model over learn?). 4. If multiple models are identified, use a mix of information-theoretic measures (e.g., Akaike Information Criterion) and end-use criteria to select an appropriate model structure. Remember we seek the best working model, not necessarily the right model! 5. Where forecasts are computed, confidence bounds should be provided. Below the list is a pink box with the text: "Any report from a data analysis exercise should include estimates, their error bounds and results from statistical tests of validation." The slide footer includes the NPTEL logo, the name "Arun K. Tangirala, IIT Madras", the course title "Introduction to Research", and the slide number "19".

Data Analysis

Assessing and reporting your results

- ▶ Compute confidence regions and errors in estimates.
- ▶ Conduct hypothesis tests on estimates.
- ▶ Cross-validate models on fresh data sets (did the model over learn?).
- ▶ If multiple models are identified, use a mix of information-theoretic measures (e.g., Akaike Information Criterion) and end-use criteria to **select an appropriate model structure**. Remember we seek the **best working model**, not necessarily the right model!
- ▶ Where forecasts are computed, confidence bounds should be provided.

Any report from a data analysis exercise should include estimates, their error bounds and results from statistical tests of validation.

 NPTEL

Arun K. Tangirala, IIT Madras

Introduction to Research

19

And finally, once you are done with all your analysis, this is the most important part - assessing what results you have obtained and reporting them is a very important step. Always subject any parameter, estimate to a hypothesis' test - whether it is significant enough, whether you have just obtained a non-zero value by chance, whether truly you should have obtained zero. So, you should subject hypothesis test of the parameter being zero – that's null hypothesis, and then, carry on in a statistically sound manner.

And secondly, report confidence regions and errors in estimates. In fact, computing confidence regions and errors are an integral part of your hypothesis test. And then, if you are building models, then cross validate models. When I say models here, need not be regression models, it could be classifiers also. And if multiple models are identified, use a mix of criteria; as I just mentioned earlier, there may not be unique answer; typically, that is the case. My students always ask me, well I have two or three models explaining the data, which one should I pick? Well, the first thing is to make sure that you followed a very systematic procedure to minimize the number of solutions that you

have and even despite the best of efforts you may end up with multiple solutions, very common. Then you apply what are known as information theoretic measures such as Akaike information criteria or Bayesian information criteria. And also keep the end use of the model; may be your model has to be implemented online; so, simple model is preferred to a complicated one with a bit of sacrifice on the accuracy of the prediction and so on.

So, again there is no unique answer, but in the end, the final results that you report should have two things: a statistical backing to it and a sound argument to why you have decided to report these results and why the results are the way they are.

So, therefore, you should remember data analysis is not just science alone. Yes, a big part of it is science, but an important part of it is also an art. Data analysis is, therefore, both a science and an art. And never report estimates without reporting confidence regions and errors; as much as possible, you should definitely do that. Of course, in very complicated parameter estimation problems, it may be very difficult to obtain confidence regions, but you should try despite a best of your... try to put in the best of your efforts to report the errors and the confidence intervals.

(Refer Slide Time: 63:07)

Data Analysis

Bibliography I

-  Bendat, J. S. and A. G. Piersol (2010). *Random Data: Analysis and Measurement Procedures*. 4th edition. New York, USA: John Wiley & Sons, Inc.
-  Johnson, R. A. (2011). *Miller and Freund's: Probability and Statistics for Engineers*. Upper Saddle River, NJ, USA: Prentice Hall.
-  Montgomery, D. C. and G. C. Runger (2011). *Applied Statistics and Probability for Engineers*. 5th edition. New York, USA: John Wiley & Sons, Inc.
-  Ogunnaike, B. A. (2010). *Random Phenomena: Fundamentals of Probability and Statistics for Engineers*. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group.



Arun K. Tangirala, IIT Madras

Introduction to Research

20

These are some of the references. Of course, no way **it's** exhaustive. I have taken some of the examples and some material **from** these books; these are quite common books of which you may find the first book on Random Data Analysis very, very useful to you because it has been written by people who have practiced for 30-40 years, seen different kinds of data, and they have examples, very good examples in their text books. **And** the remaining three, at least the book by Johnson and Montgomery and Runger, they **they** are classical books in probability and statistics - applied probability and statistics. The book by Ogunnaike gives you the fundamentals of probability and statistics for engineers. If you are an engineer, you may be really interested in looking at the book. Of course, there are hundreds of other books that you can refer.

So, hopefully you enjoyed the lecture and you understood at least what is data analysis all about, what precautionary steps that you have to take, and how careful you have to be in data analysis, what is the systematic procedure, and some terminology. We will quickly look at a couple of illustrative examples in R and that will hopefully supplement this lecture.

Thank you.