

**Applied Time-Series Analysis**  
**Prof. Arun K. Tangirala**  
**Department of Chemical Engineering**  
**Indian Institute of Technology, Madras**

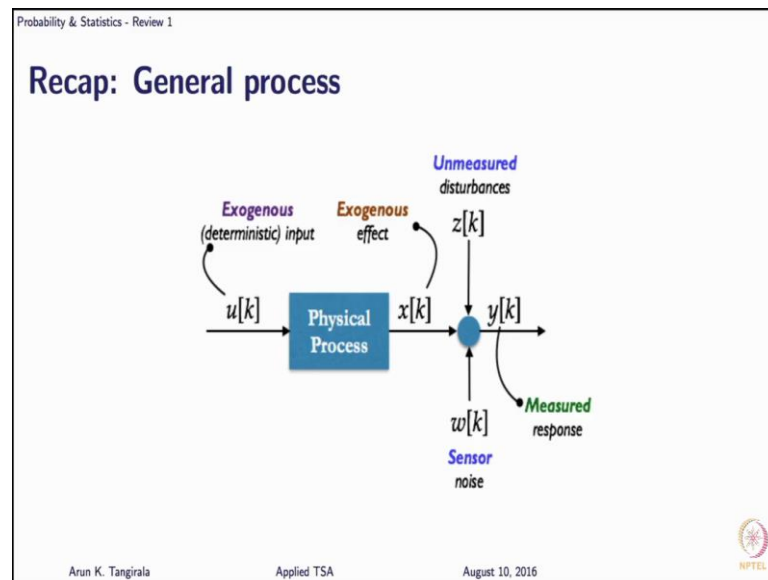
**Lecture- 07**  
**Lecture 04A - Probability and Statistics Review (Part 1)-1**

Today we will formally begin our course on time series analysis. What we will do today is will now start looking at formal concepts of that are necessary to do a proper analysis of the time series data. It is going to be a lot of theory, I should caution you up front, but this theory is an inevitable, I would say they will I do not like using that term, but it is just inevitable, I mean you cannot get bored with it. We will try to make it interesting from time to time, we will try to show you examples; sometimes in our, sometimes just worked out examples by hand and ultimately all of this is going to play a significant role in how you analyze the data, how you treat your data.

Therefore you have to pay attention to the theory because in every data analysis there are going to be assumptions and if you are not aware of those assumptions then it is likely that you will choose a wrong techniques for analysis or you will end up doing a complete mess of the analysis as I call it as dialysis, but you have to be really careful and many a times your analysis in the first round can give you probably not in the information that you are seeking for but that is where that is where the difference between a learner and a blind user comes into play. At that point in time you would be able to diagnose after going through a formal course and time series analysis, you would be able to kind of guess as to where things could have gone wrong and sometimes you may have to actually question the assumptions that you have made, whether they are valid for the data and so on.

So, let us begin our journey and before we do that let us actually recap what we have learnt in the last week so that you are clear in what framework we are going to work in.

(Refer Slide Time: 02:16)



So, what you see on your screen is a schematic of a general process; I have spoken about this in the last lecture. Any process that you would encounter more or less can be represented in this way of course, there are some assumptions here and the main assumption is that whatever effects of unmeasured disturbances and the effects of sensor noise that you have are going to add on to the response of the so called physical process, what we mean by physical process is a process that you are able to kind of touch or see like maybe a spring mass system or an atmospheric process or whatever may be a biological process and so on.

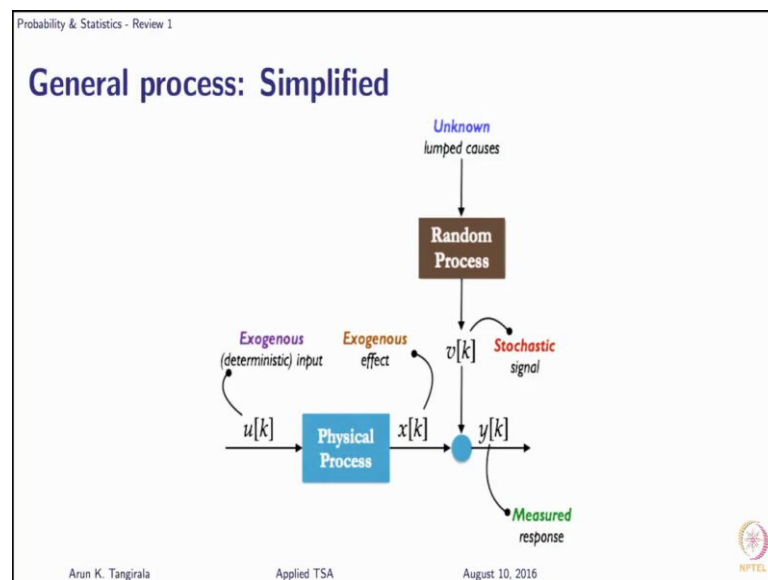
So, as engineers we are generally concerned about the physical process that is we would like to have a model of it and the typical approach is to look at the so called input, output data. The input here is being represented by  $u$ ; you can call it also as a cost and then there is a response  $x$ , based on this input  $u$  and response  $x$  we would generally try to build a model. Unfortunately we do not have access to  $x$  because we would use a sensor and on top of it there would be effects of uncontrolled or unmeasured disturbances. So that ultimately what you have with you is what is being denoted by  $y$ , I will talk about the notation a bit later on in detail but right now they are fairly self explanatory.

So, what one would have in general is the input  $u$  and the output  $y$ . In time series analysis a scenario is slightly different in fact, quite different; I will come to that shortly. So this is a general process description that one would run into, where we have made this

important assumption that the effects of unmeasured disturbances and the sensor noise add on to the response of the physical process and as I say always say this is the assumption on which a lot of research has been done. And once we are saturated with this assumption that is we have managed to explain a class of processes with this additive assumption and then we encountered those processes where this additive assumption does not work. Then we turn to multiplicative assumption and that will allow people to carry few generations to carry out research, explain a few more class of processes and all of that is done then you can think of all other square root, division and so on, we will not be there perhaps to see that, but let us worry ourselves about the with the additive assumption and that is what we will focus on.

Now, what happens is in general we will not be in a position to segregate the effects of unmeasured disturbance and sensor noise. So we lump all of this into a single effect or a single cause; you can say. So, in this fashion here and also say that I am unable to explain the effects of unmeasured disturbance and sensor noise using a deterministic model, what we mean by it is a mathematical function.

(Refer Slide Time: 05:38)



So the difference between the schematic that you just saw earlier and what you see right now is the lump nature of the uncontrolled or the unmeasured causes and the sensor noise as well. We have lumped everything; we say we do not know what is causing of

course, there is also possibility that you can have effects of measure disturbances; that can always be brought in, we do not have to worry about that at the moment.

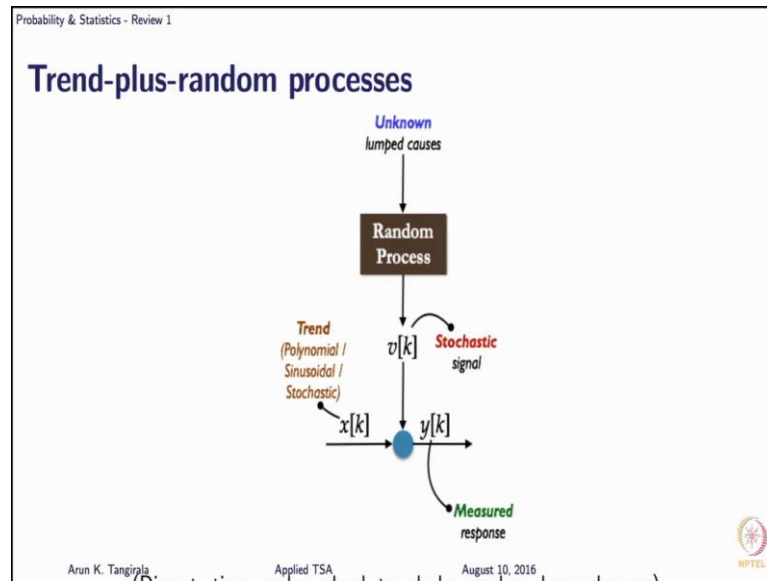
So, we have lumped those effects that you saw earlier into some unknown causes and we believe that these unknown causes are driving some random process which is in turn producing this lumped effect. All of this is imagination, we are actually doing a lot of abstraction here, but this is an organized friction, it is friction but it is an organized friction which helps, which works. So this is how we are imagining the process to be wired within, it is not necessary that the real process is actually wired this way. We have no idea how the real process is wired, in general that is the fact of life, but what we are seeking is some kind of abstraction that helps us in making predictions or in classifying or in detecting periodicities and so on.

Now this is a simplified version of what you have seen, further now when we move into time series analysis at this stage what you are seeing is the subject of system identification, where the input  $u$  is given and the response  $y$  is also given to you and the goal is to get a model of the physical process. Now, that is a general scenario that we all run into; however, if you look at the history of system identification time series analysis precedes system identification, where people said well I am just given some set of observations of a process; can I predict, can I make a prediction of what happens next.

In which case, there is no attention paid to the causes; it does not mean during those days people did not know something like a cause effect model existed and so on, that is not the case at all. It was necessary to address a problem; for many reasons right because in many situations the causes are not known and in situations where causes are known, this may be the scenario where the causes are known, but the known causes are unable to explain everything. So, you will still have to explain a part of the effect using some other model and that model is what we call as a time series model.

So in other words if you actually look at this schematic here, there is always going to be these effects of unmeasured disturbances and sensor noise. Even if you do not have unmeasured disturbances, you will invariably have sensor noise corrupting your data. So, there is no escape to that channel which is being labeled as random processes.

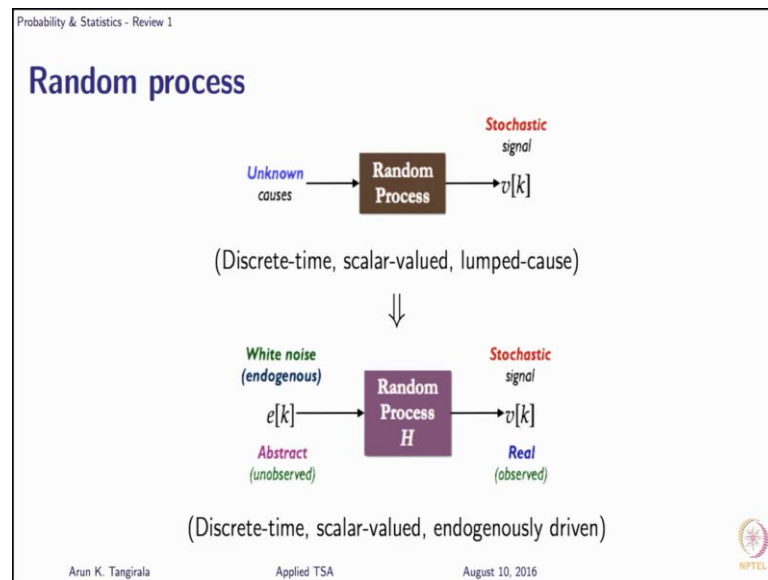
(Refer Slide Time: 08:43)



So, now we move from system identification framework to the time series framework alright where we have now thrown away the costs. We have kept aside the causes not because we do not know them but we will focus in this course entirely on this part alone and when we are well versed with this and you are not fed up with time series analysis or you are not fed up with the instructor then you can take system identification and then learn how to do is wireless program.

So, what we have done is we have kept aside the causes assuming that there is a not known or even if they are known I cannot measure them and this is the scope of time series analysis where I have signal  $y$ , which is made up of two components; some  $x$  which is coming out of some physical process I do not know what is generating  $x$ . In the time series language this  $x$  is typically known as a deterministic trend or sometimes a stochastic trend also but the general generic term for  $x$  is a trend and this trend can be either explained by a mathematical function or you can make it a bit more complicated and say no  $x$  is itself coming out of some random process, we do not know; we will consider those situations a bit later, but what we have done is we have thrown away the  $u$  and now we are imagining that sorry  $y$  is made up of  $x$  and  $v$ .

(Refer Slide Time: 10:27)



We will even simplify this and come up with this random process where we have thrown away  $x$  also. Now we are saying that there is this  $v$  that I am measuring, not  $y$  as such; well these are thus notations what I mean to say is that we have thrown away the trend part, we have just looking at the random part of this measurement and saying that I have access to  $v$ , which I am imagining to be driven by some unknown causes that is exciting a random process. So there is an imagination again here, if throughout the course we are just imagining and it is a mathematical plus statistical abstraction.

So, this process that you see on the top is a very simplified way of looking at a random process where unknown causes are driving the random process and producing  $v$  which we call as a stochastic signal or the random signal. There is a subtle difference between random and the term stochastic, but will not observe the difference. What we have now is  $v$  that is observations of the signal  $v$  and what we do not have is these observations of unknown causes and the goal is to build a model for the random process or you can say the goal is to understand, draw inferences about this random process.

Now, we make a further abstraction of this and actually come up with a slightly different version of this process where we replace the unknown causes with some unknown signal. So this unknown causes are actually lumped, they are all you can say a net set of causes from unmeasured disturbances and sensor noise and sometimes even modeling errors.

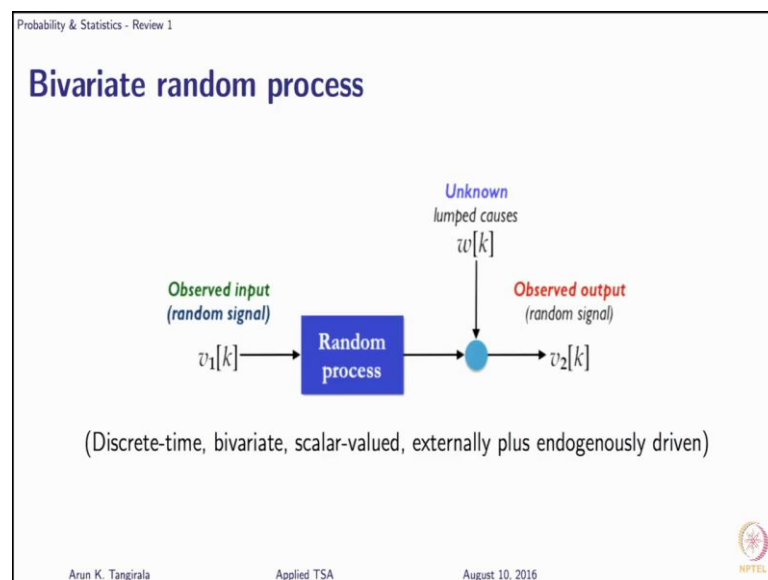
Now, it is hard to actually build a model without imagining an input and therefore, it is convenient to introduce a fictitious input that is driving this random process which we call as white noise all of this will become clear later on, but I am just giving you a holistic picture of what you are going to see in this course or what you generally see in time series analysis. So, to summarize we start from a general description where of any process where there is a physical part or a deterministic part and then there is a random part and we keep aside the causes either because we do not know or even if we know there is anyway need to explain the random part, and we lumped all the unmeasured causes and maybe the sensor noise everything into a single cause and assume that whatever random signal that I am observing is being driven by this lumped causes and then finally, say that I am going to replace this unknown causes by some fictitious input signal called white noise.

The jump from the top schematic to the bottom one is huge, it is huge and we will learn in due course of time what allows us to actually represent a random process the way we have done in the bottom picture where there is a white noise driving the random process and also I mean producing the random signal that we see. Actually lot of effort went into formalizing these ideas, deriving conditions under which it is possible to write such a representation; that means, it is not necessary that for any given random process I can write or I can represent the way I have done in the bottom schematic where there is a white noise driving a random process. In fact, that random process is a linear random process; it will restrict ourselves to that.

So there are mathematical conditions, some statistical requirements on the random process that will only when those are satisfied, you will be able to actually represent a signal that you have observed as white noise passing through some linear random process and will denote this random process by  $h$ . So, I have given clearly the framework for you and then will proceed to the formalization of ideas, so any questions on this; as I said it will take some time for us to understand under what conditions we can represent a random process the way we have done that is white noise exciting a linear random process  $h$  and producing the signal of interest alright that will require both understanding of the time domain statistical properties and as well as a frequency domain properties of the random process.

We will assume that whatever process we are looking at, that we are going to examine; that it meets the requirements of being able to be represented as the way we have done. So, sometimes we will also digress to a bi-variate random process, what we have been discussing until now is a scalar valued signal which means you are looking at a univariate case but occasionally we may move on to the bivariate random process case where now the schematic looks pretty much similar to what we started off with but the difference between the first schematic and the one that you see on the current slide is that what is the difference, input is random and secondly the physical process has been replaced by a random process.

(Refer Slide Time: 16:11)



But you can say that in fact that need not be a restriction, the random process that you see there could be a physical process, but the main difference is that the input is also random. What do you mean by input being random, which means that that input also has some uncertainty in it whereas, in the first schematic that I saw the input is deterministic, I am giving the input I know exactly what the input is; here the input also has some uncertainty.

Now what can in what kind of situations do we run into this scenario, well whenever you measure input you are not actually controlling the input a classic example is let us say you are building a model for relative humidity in terms of temperature, we have talked about that example earlier. So, the output is relative humidity and the driving signal is



temperature, and let us say we are looking at some geographical region. We do not have control over temperature, can we actually perform some experiments in the atmosphere where I say; I will change the atmospheric temperature let me actually find out what the relative humidity is; that is not possible isn't it.

So what we do is; we observe the changes in temperature, we observe the changes in relative humidity and physics tells us that there is a close connection between temperature and relative humidity. Therefore, instead of building a univariate kind of model for relative humidity, I can probably make a better prediction if I know the temperature also right. If I were to make a prediction of relative humidity, I have two choices; I only rely on the relative humidity data on its history essentially and build a univariate time series model, that is one option and the other option is given temperature data, if I am given that there is that the history of temperature series then I can make use of that and now build a model between temperature and relative humidity. In other words I will use the past of temperature also in making a prediction of relative humidity, intuitively we should expect better predictions.

Now having said that; temperature alone may not be the factor that will help me predict  $r_h$  accurately, there may be other factors contributing to relative humidity; pretty much like what we said in the first slide. It may not be possible to explain the changes in relative humidity using temperature along which means there are going to be some unexplained effects in the relative humidity, either it could be due to sensor noise or maybe other physical variables which we may not have measured. So, what we need to do therefore, is still account for those unexplained effects through another channel and we lump all those unknown causes and once again we build a model.

Now, in this case we build a model; we build two models, one model relating the temperature to relative humidity that is a signal  $v_1$  to  $v_2$  and in addition a model for the unknown effects, unexplained effects that would be again a time series model. So, you can see at the heart of all of these analysis is a univariate time series model.

One needs to be first well versed with building a time series model given just a signal without the causes and then you can take all that knowledge and embedded either in this kind of a system identification scheme, this is also system identification; this particular class of problems in system identification are called Errors In Variables case; EIV case

where as a classical system identification assumes that the input is known accurately when which means input is free of error. The errors in variables class of problems assume that input is also known with error only with uncertainty, but in both cases we need to be well versed with building time series models. So, we will therefore, for most part of the time in this course, confine ourselves to learning how to build time series models, given a signal only the history of the signal; how to make a prediction.

(Refer Slide Time: 20:59)

Probability & Statistics - Review 1

## Framework

1. Univariate / bivariate
2. Linear random process
3. Stationary and non-stationarities (of certain types)
4. Discrete-time
5. Time- and frequency-domain analysis

**The cornerstone of theory of random processes is the concept of a random variable and the associated probability theory.**

Arun K. Tangirala Applied TSA August 10, 2016 NPTEL

So, the framework in which we are going to learn this is of course, univariate; occasionally bivariate as I mentioned. We are going to look at linear random processes and further we will assume that processes are stationary, but we will also learn how to handle certain types of non stationarities and we will not talk about that at this moment. When I talk of stationarity I will mention that and then as I have said in the last week, we will restrict ourselves to discrete time random processes and equip ourselves with the tools in time and frequency domain analysis.

Now, the cornerstone of the theory of random processes is the theory of random variables that is the base, if you are not familiar with the theory of random variables or in other words probability and statistics then you are going to find this course difficult and that is the reason I have asked each of you to sit through that short 10 to 12 hour course on introduction to statistical hypothesis testing which contains; essentially will give you all the background that is required. Nevertheless, I will go through not a good strain kind

of review of the theory of random variables and maybe Shatabdi not get them on kind of review but we do that; but if you have questions you should ask.


(Refer Slide Time: 22:26)

Probability & Statistics - Review 1

### Notation

- Random variable: UPPERCASE e.g.,  $X$ ; Outcomes: lowercase e.g.,  $x$ .
- Probability distribution and density functions:  $F(x)$  and  $f(x)$ , respectively.
- Scalars: lowercase  $x, \theta$ , etc.
- Vectors: lowercase **bold faced** e.g.,  $\mathbf{x}, \mathbf{v}, \boldsymbol{\theta}$ , etc.
- Matrices: Uppercase **bold faced**  $\mathbf{A}, \mathbf{X}$ .
- Expectation operator:  $E(\cdot)$
- Discrete-time random signal and process:  $v[k]$  (or  $\{v[k]\}$ ) (scalar-valued)
- White-noise:  $e[k]$
- Backward / forward shift-operator:  $q^{-1}$  and  $q$  s.t.  $q^{-1}v[k] = v[k - 1]$ .
- Angular and cyclic frequencies:  $\omega$  and  $f$ , respectively.
- ...

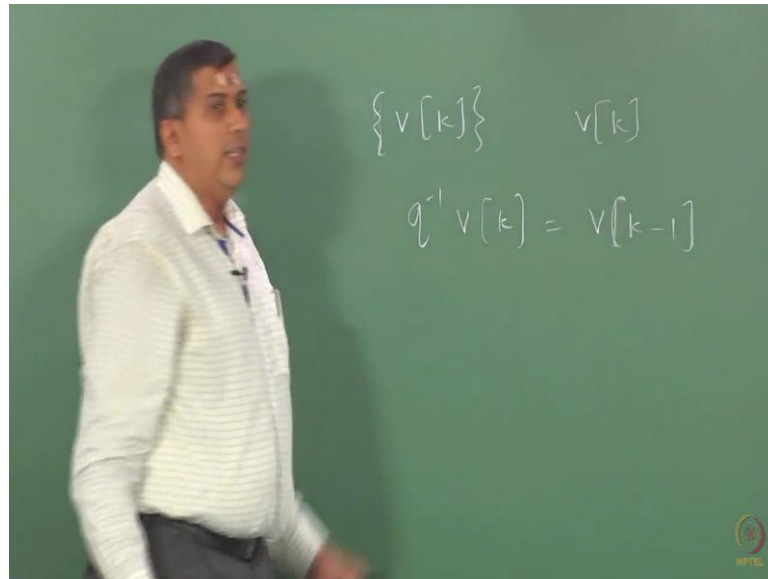
Arun K. Tangirala Applied TSA August 10, 2016



So, we will introduce a notation first and then will get into the review of theory of random variables probability and so on. So, all random variables as you know are denoted conventionally by uppercase and the outcomes of the values that they take by the lowercase variables and probability distributions and density functions like likewise are denoted by uppercase and lowercase  $f$  and scalars throughout the course will be denoted by lower case. Whereas, vectors will be also denoted by lower case but bold faced, on the board it may be very difficult for me to draw a bold face, but I will indicate that, may be all that you have to do is keep up a bold face while going through that and then you have matrices which we will denote by uppercase bold faced once again.

Very often you will run into expectation operator, you are really expected to be well versed with this operator; that is one operator that you should be very comfortable with and discrete time random signals and processes will be denoted as either  $v[k]$  or in curly braces.

(Refer Slide Time: 23:48)



Typically, if you see many texts you would see a random process being denoted with  $v$  of  $k$  in curly braces but many a times in problem statements, you would say consider a random process  $v k$ . Now this  $v k$  is actually used to denote a random signal, a random process and at times the observation of the random process at the  $k$ th instant. So, you should basically be able to interpret this based on the context and I think people who speak Tamil for example, should be quite comfortable because you have four sounds for a single letter.

So, Gandhi can be county if you read wrongly alright, so based on the context you should be able to interpret and then of course you will encounter white noise, I mean it will keep making noise throughout the course, so you should be really comfortable; everywhere white noise will be denoted by  $e k$  unless otherwise specified and we will also frequently make use of the backward and forward shift operator. The backward shift operators will is denoted by  $q$  inverse, its role is to shift an observation at let us say  $k$ th instant to 1 observation in the past.

So,  $v$  at  $k$  minus 1 is observation of the random signal at  $k$  minus oneth instant,  $q$  inverse is an operator not a multiplier and  $q$  is a forward shift operator; you should be able to guess what it does. So, when you encounter  $q$  inverse unless otherwise specified it is an operator, it is a backward shift operator and we will run into angular and cyclic frequencies when we dwell on frequency domain analysis. As usual we have the standard

symbols  $\omega$  and  $f$  with the appropriate units and there will be other notations that will introduce, which will be mentioned as the need arises.