Dealing with Materials Data: Collection, Analysis and Interpretation

Professor M P Gururajan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay

Lecture No. 99
Course summary

Professor Hina Gokhale: Hello and welcome back to the course on Dealing with Materials Data. I am Hina Gokhale.

Professor Gururajan: I am Gururajan.

Professor Hina Gokhale: And we both have taken a journey along with all of you through the process of learning the basic sub statistics and applying and using the tools to understand what the statistics does to the data. The purpose when we started this course was very simple that we wanted to have basic understanding of statistics before actually the tools are been put to use. So that the analysis that one does with the tools becomes meaningful.

Also the purpose was the before, in some cases, before you perform the experiment itself, if you do a little bit of a planning for experiment and understand statistically as to what analysis that you wish to do in future and for that what kind of you know experiment should be designed that also can come handful, handy to you. But as I said before, all this basic understanding is insufficient until you have the tool in your hands and I am sure Guru will agree with me on that front.

Professor Gururajan: Yes, surely. So the tool that we have used for this course is R, which is a open source programming language, it is freely available and you can use it both in Linux and Windows and Mac, whichever machine that you use, so it does not matter which platform you use and on top of it, there are lots of libraries that are freely available and its very active community.

So as we have seen in the design of experiments session, there are programmes that are being made available by people that are being developed and it also allows us to actually contribute to this community if we come up with a new library or program or any such development that we make. So that way it is a useful tool you have and it is a very nice tool to use and I hope you had enjoyed working with our for the past 12 weeks, 12 weeks, yeah, so 3 months yeah.

Professor Hina Gokhale: Yeah, it is a long time.

Professor Gururajan: Long time, yeah.

Professor Hina Gokhale: Well, I would like to repeat once again few things here. You know, these days, the computing tools and computational techniques are very easily available and accessible to us on our even small laptop and they are very attractive because there goes in a data and here come out a beautiful graph or nice table explaining something. The question is have we used the right tool?

Or the question is whatever graphs and tables have come, what do they say? Also the question becomes is that there are many options available, which option to choose? These are the questions which can be answered only by understanding the basics of statistics. I only again say that if you want to use machine learning, if you want to use artificial intelligence, it is important to know the basics of statistics.

You do not have to learn the great theories of statistics and probability, but if the basic, you know concepts are clear, if the assumptions are known, you know that when you try to use these tool what errors can occur or is this a correct tool for this purpose? If this information is available with you, then the tools are put to use in the better way and therefore this course has throughout emphasized that what is the basics of statistics?

What are the assumptions made, from where did this procedure come and how to interpret the results and therefore we chose the open source software, so that it is not bound by you know what you buy or what your organisation to afford to buy, but you can use it at anywhere and start becoming almost a data analyst, what do you say, Guru?

Professor Gururajan: Yeah, surely. But like everything else in life I think we need a balance here.

Professor Hina Gokhale: Yes.

Professor Gururajan: Yeah So, my advisor used to say that nobody gets drenched in a simulated range so you can freely go, take the computer, do lots of things with simulation and coding and programming and things like that. So that is one viewpoint, there is merit to it, I also strongly urge you to play with R, take the data, do things, some of them might be silly, some of them might be stupid, some of them might be wrong and some of them might be great, so its fine to do that.

One the other hand I have also seen textbook on Numerical Methods or Differential equations which said that you should resist the temptation for computation as long as possible. In the sense that you should not just jump in because you just have a tool, you know, I have a hammer so everything is a nail, I am going to keep banging on top of everything, so that is not the right way, so there is a time to play, there is time to explore.

But there is also, and there is lots of learning that happen when you are relaxed like that and you are just fooling around with things, but there is also time when you have to know the basics, the background, the setting in which we are doing what we are doing ad intelligently steering our calculations through when we phrase difficulties or when we have some results interpreting it and doing the next right thing.

So that way it is very important to pay attention to the basics which is the statistical theories in this case and them using it and implementing it in R. And so when you reinforce these two things, the balance also brings you I think much deeper understanding.

Professor Hina Gokhale: Yes, I agree. Because I also find that just learning the theory for one thing it becomes very boring, it does not remain any more interesting, so if you apply the tools then you come to a stage where you realise, you actually see the realisation of what theory you have learned, sometimes you learn by mistake also. You have a hunch as they state in art of data science that you have to have a hunch.

It is, you cannot be exploring absolutely in dark using statistics. So you have to have a hunch that my answer has to be in this vicinity, or the solution has to come something of this form, when it does not come, you have to realise that something has gone wrong somewhere, now it could be the tool and it could be the understanding of the theory, so both of them go hand in hand.

And I think that is why we made it the point that this course we go through by having a theory and a practice. A theory and a practice so that things could match with each other and you also do not get tired with this statistical theory, so one week of theory and one week of playing around with data is what we have designed and I believe that it has worked well.

Professor Gururajan: So we have also made sure that even though things are together but they can also be independent by themselves, so there is an option for you to just do the odd weeks or even weeks, the theory weeks or the practical weeks. So you can do one, but it is the maximum benefit that you can derive is when you do both, but it will not necessarily have to alternate, so you can finish one and you can do the other or you can keep going back and forth which is most recommended thing from us.

Because you do something, you go deeper in to theory and then go take a practical case, take a data, try to do all that that you have learned, think about it and like you know I was saying, you should know what is the number that you are expecting. Know I had a physics teacher that said, "Before writing down approximately what you are expecting, you are not supposed to solve the problem," and once when you do this, you also develop an intuition for the number.

You also develop an understanding for things, you know, you expected to be this and the surprisingly you get something else and then you understand why and that is where the real understanding starts coming in. So I have seen that with data you have an innate understanding which I do not still have but that is because of continuous working like this and thinking through things and so we hope that you do whichever fashion you like but both the parts so that you will get the complete picture.

Professor Hina Gokhale: Yes and with respect to statistics also we have not kept it very heavy with theory as you must have noticed, very-very basic concepts with starting with descriptive statistics and probability and random variable and some special distribution and we got into immediately with the estimation and hypothesis testing, but the real applications have started with regression in which we have not gone into much of a theory.

And in this area, I would like to bring it to your notice that you must be feeling that on one hand somewhere in the course we have said that in the field of material science and material engineering, the data tends to be not in a beautiful bell shaped curve, they generally tend to be skewed and then we use central limit theorem and then we try to, you know, develop the theory for hypothesis testing, regression etc., using only normal assumption.

But if you recall, when we go to the ANOVA, it is made very clear that you do not have to stick to the distributions. The T statistics, the chi square statistics and the F statistics, the Z statistic will hardly play a role, it will never happen that you already know a variance, so I am discounting that, but this three statistic, basically you have to understand is that if the values comparatively look very large, then it is not meeting your hypothesis.

It is not meeting with your understanding that some parameter is zero and therefore it is a critical region. It is true with respect to the F statistic, it is true with respect to T statistic only when you want to look into those tables or exactly find out what is the probability of critical region alpha, type 1 error, you need to make a assumption on distribution, that too I would say if you are doing simulation.

What I would do is I would simulating about 10,000 times and find out the right number where does it fall by you know rearranging them in an order like we did for the Weibull distribution you do not have to actually know the distribution at all, you just generate the number and do it.

So the point here I wanted to make is that if anytime you fail that on one hand we are learning everything with the assumption of normality but on the other hand Weibull and log normal are more important for us is what is being said and the you have on skewed distributions then the solution lies here that it is the F statistic, T statistic per se, not the distribution which are important.

Professor Gururajan: So from the point of view of using R, what we have seen is that you can get lots of information just by plotting. You should try to explode the data as much as possible and visually reading information is still very useful and it is a good skill also to have, so that is one. Second thing you should no hesitate to do simulations if it is needed and so we have dealt with simulation in several different scenarios.

And we have done different kinds of simulations in a boot strapping or calculating or estimating pi, for example, so different kinds of calculations that you can do using simulations. In general, I have found that students typically depending on what they do get some kind of mental block, the students who do more computational type of work, typically spend more time of physics based models, solving PDEs and things like that.

Statistical simulations are relatively rare. On the other hand those who do experiments do not even look at statistics or do any analysis, even basic ones that you can do simply by taking origin or excel or R for that matter. So we wanted to bridge this gap, students who are comfortable with the terminology, who are not put off by terminologies like hypothesis testing or Bayesian inference.

So just to give them that one familiarity that they are not afraid of hearing these words and it make some sense to them and the second thing after they become familiar, they try to explore it on their own and try to do similar things on their own. So that is the purpose of this course that they are familiar with the terminology, they know what is the, at the backend and they are able to themselves using R.

Professor Hina Gokhale: Yeah, I think overall that was our purpose and I think see this metallurgy has a, in the eyes of statistician I find that it is a very peculiar subject, maybe it is there in others, I do not know. Generally the data generated generation is very expensive, so you have two or three data at the most five data and we are very joyful and celebration mood for that that we have got a 5 or 7 data points.

On the other hand today materials data science is going to play a major role because a lot of past data is being generated and it is being stored as a database and we would like to make use of it. So, I feel that material science and materials engineering need very strong techniques of statistics which, in which on one hand n tends to infinity and on the other hand n is very small.

And this covers the complete span of statistics and computation only makes it possible. So boot strapping technique as you said or you know many other simulation techniques are the only ones in our hands when the data is very small, why? When the data is very large the Bayesian inference, regression models, etc. etc. You can go further, even beyond this course, you can use heuristic models such as artificial neural network or genetic algorithm or you know any of this heuristic models and use it.

But when I say this gain this course plays a very important role because when you use any of these models particularly I am referring to the heuristic model, the type 1 and type 2 error play a decisive role in deciding how good is your model, and sometimes it is forgotten and it need not be forgotten, it need not be forgotten, I can go ahead and say that you should not be forgotten because that gives the strength to your model.

And that is one another take away I would say from this course that even if you try to apply data analysis techniques beyond this course, some basic concepts that have been given in this course may please be kept in mind, because here they may have been given in the context of, for example, what I gave right now is an hypothesis testing but then it applies to any situation where you are trying to make a judgement using a data and fitting a model to it.

Professor Gururajan: Yeah, so to go back to the original point about the data. So existing data is very limited, so one of the things that we expect that will happen after you do this course is that you will curate data better, even when data is available, for example, we discussed it in the case study on Hall-Petch. Many information about the data is missing, for example, there is a grain size, but the distribution information is missing.

So unless you have the complete information, then making inference based on that data becomes problematic and so we hope that all of you irrespective whether you are doing experiments R modelling will curate you data better. Second thing you will also make it available as much as possible for everybody.

Mostly data is analysed and it is presented in a form where the original data is lost and sometimes if you want to redo some of the analysis, it becomes important to have the data in the original form. So we also hope that you will get into the habit of curating the data better and sharing it as widely as possible and go back to the literature, again going back to the Hall-Petch study.

For 70 years people have been using Hall-Petch and the conclusion we drew is that there is no evidence neither experimental nor theoretical to justify Hall-Petch. So maybe there are surprises waiting there if you go back to the data in the literature and put it in a form in which you can analyse it at a much broader level. So we hope that this course will help you do both and you will do it and get back to us about your experience.

Professor Hina Gokhale: So overall can we say Guru that this course is a beginning.

Professor Gururajan: Yes.

Professor Hina Gokhale: It is a it is just a beginning, you are starting a journey and we invite you to explore either side, you can explore the side of complete computational data analysis or you can also explore the side to learn more details of a say machine learning and in machine learning different techniques that are being used and how the statistics applies there and how you can improve upon your computational capabilities by using those theory.

So I think this course is just a beginning, we cannot even call an introduction, but it is just truly a dealing with materials data as a first step you know and you get the feel for the data is where we have started.

Professor Gururajan: Yeah, and we will also give them an open invite that we both are open, anytime they explore, they want to find something or they have found something new, they are welcome to get in touch with us.

Professor Hina Gokhale: Yes and we encourage you that you further explore the field of a materials data analysis because it is going to be one of the very important field in the future, so we wish you that you this is beginning of a very long journey, a fruitful journey and a satisfying journey.

Professor Gururajan: So welcome to materials data and have fun.

Professor Hina Gokhale: Thank you.