Interactomics: Basics and Applications

Prof. Sanjeeva Srivastava
Prof. Santosh Noronha
Department of Biosciences and Bioengineering
Department of Chemical Engineering
Indian Institute of Technology, Bombay

Lecture – 54
Basics of Statistical Analysis – I

Today we have a guest lecture by Professor Santosh Noronha; he is a faculty at IIT Bombay. Who works in the areas of big data analysis and various device development. Today lecture is about data analysis in evaluating the various hypothesis like searching for a potential drug delivery target or a genetic target in a disease context. When we are trying to search for a biomarker candidate or a potential drug candidate, it becomes very crucial to know how to make a good experimental design in which manner our data should be reproducible.

If you think about a test which is has to be given to the patient, the test has to be reliable reproducible as well as it has to serve the large community. So, therefore, your number of patients to be included for the study has to be much larger. Your study design has to think about various flaws various mistakes various errors which might be happening if you do not carefully consider your end goals, your actual questions to address.

So, in this light Professor Noronha is going to illuminate your understanding about how one should think about the good experimental design before actually planning and executing the omic experiments. By now you are familiar that data generation using protein microarrays or using various type of NGS platforms or various type of proteomic technologies are quite straightforward. As long as your sample preparation is good as long as you know what exactly biological question you want to ask for.

However to really get a meaningful data meaningful information it is not so straightforward. And that is where you need people who are good knowledge for a statistics, who can work with you in designing good experimental plan. So, if the aim is to look for biomarkers or to choose the drug target your experimental design your study plan your number of patients replicates all of these things becomes very crucial. So, let us welcome Professor Noronha to give you this lecture about how to really carefully plan a good experimental design for omic based studies.

From what I see as a background of most of the participants here dear students that too from a life sciences background most of you, there is some researchers here as well I understand. So, I thought it would be very useful and provocative both to bring up a discussion of data analysis as it applies to multivariate experiments such as the ones you are doing in proteomics.

So, a big issue with statistics is how confident are you about any insight you get out of any analysis that you do. And while clearly you are spending 3 to 4 days learning about experimental protocols, the question is what are you going to do with this data that you generate and what hypothesis does it drive and for that matter how valuable is this data that you are generating in the first place.

So, I am not actually going to take a case study and the reason I am not taking any case studies, because I am not in this domain as a core domain I am a biochemical engineer. I am interested in production of pharmaceuticals I usually do not ask questions about what is the role of this gene in that particular disease context and so on. But one of the things as an engineer you forced to do is to deal with large amounts of data especially in a process analysis context.

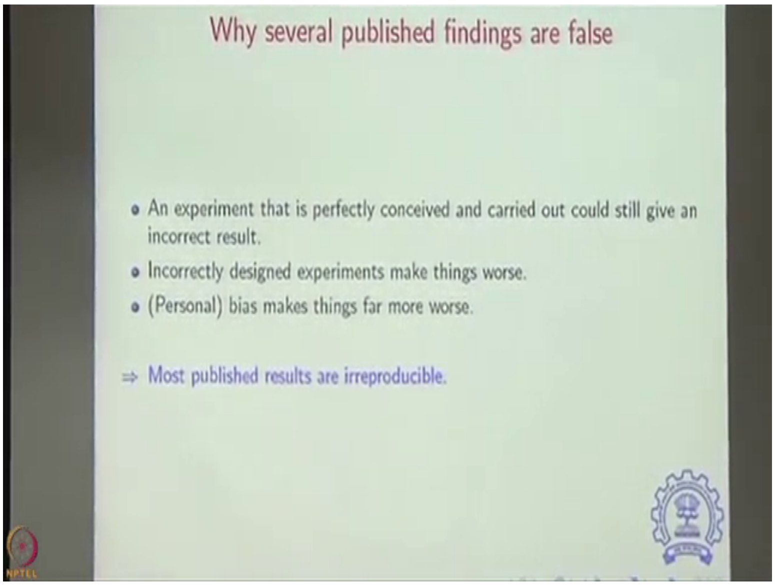
And this data analytics kind of need actually we find is very useful in evaluating hypothesis which is what most scientists are concerned with. So, these are a collection of thoughts that I have had over while which I think apply mostly to scientists less so to engineers. But they have to do with the fact that we are generating data at very large scales we are generating data in very large amounts. And the question therefore, is given this rate at which we are generating data and the assumption by the way is that experiments we are doing. A well designed experiments which are therefore, generating good data.

But given this data the question is what can you infer and what are potential errors that one can make in trying to arrive at conclusions. So, that is why I am titling this particular talk as one on reproducibility in particular. And one flipside to this is as we see a lot of publications come out especially in the omics to domains, with question comes up which aspects of these are reproducible and if they are not reproducible what kinds of mistakes are people making which prevents good insight from ok; potentially being you leveraged into some kind of future drug development pipeline.

So, remember most of the reasons for working in this space of omics is that you are trying to find targets typically for drug delivery ok. So, at every level here you can be asking the question, how I statistically found a good target second given a target how I statistically found a good drug candidate given that there is a library of a zillion candidates have you found a good candidate to deploy as a potential drug. And if you think about what is actually occurred out there the success rates are very low and success rates invariably are low because of issues to do with reproducibility and that is what I want to deal with.

If you look broadly at why most published findings are false you can break it down into different possibilities, which is that you have actually thought of a good research design for your experiment. But given the fact that for example, if you are talking about disease conditions; given the fact that you might not be able to access that particular tissue type in more than one patient.

(Refer Slide Time: 06:25)



Why several published findings are false

- An experiment that is perfectly conceived and carried out could still give an incorrect result.
- Incorrectly designed experiments make things worse.
- (Personal) bias makes things far more worse.

⇒ Most published results are irreproducible.

Therefore you are not in a position to do replicates of this as a study systematically over a larger population. You might therefore, end up which your bad luck and with an incorrect result. It is kind of like saying you will toss a coin 100 times and with your bad luck the 1 time you toss the coin 100 times you might have ended up with, 30 heads out of 100 which is theoretically possible. And 30 out of 100 sounds as if you have a biased coin not a fair coin nothing wrong with getting 30 out of 10 it is just that it is a random event, it has to do with the fact that coin tosses are random events. And with your sheer bad luck the 1 time you did it you could have ended up with 30.

Now, your problem is you are trying to ask the question why is this coin behaving in this particular fashion you would have expected a fair coin. Therefore, you would have expected 50 heads out of 100 and instead when you see a 30 you are actually pushed into this question as to whether what you are seeing is an example of an unluck event with the with the fair coin where you should have expected 50 heads, but instead you shot 30 or have you instead seen a biased coin.

So, which way is this. So, in another words are you looking for a particular hypothesis and you are saying that the hypothesis is not true, are you saying that an alternate hypothesis is true; which way do we go? So, this is a situation which instead of that fair coin now and biased coin think of whether you are looking at a genetic target for manipulation in a disease context. So, is that target something that you pulled out as a random event or is it. So, significant to you that you feel that you should now go to follow up experiments.

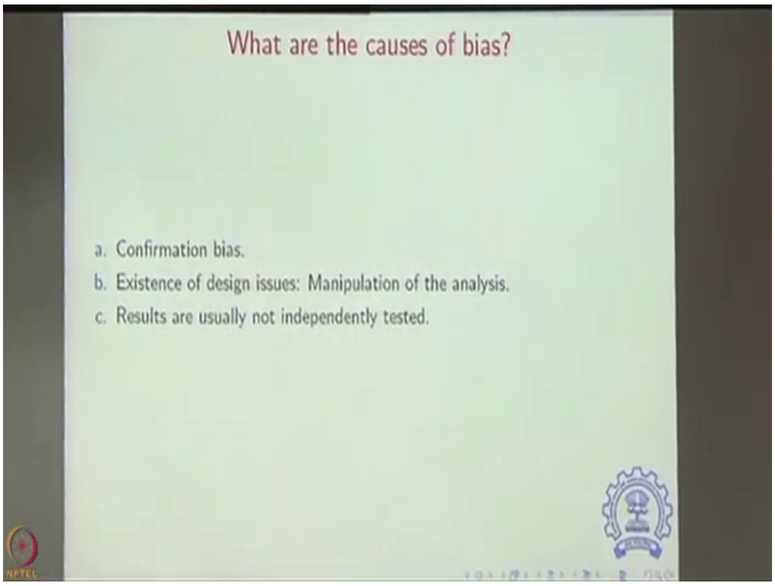
A context of this is therefore, that if you are trying to find targets and there is an investment by the way, in both the experiment that you are performed to find targets and an investment subsequently in trying to validate these targets ok. How confident are you in carrying out these experiments and of course, the assumption in all of this that the experiment is designed correctly that you have done it with appropriate controls that you have randomized that you have got appropriate ok, controlled experiments that you are doing.

And if you have not done that then things only going to get worse in terms of the quality of data that you have, and even more critically and this is a sad fact. If you as an investigator already biased towards one cluster of genes as being important to you then personal bias makes things even worse and the brute truth of this is that most published results are therefore, irrepressible.

So, much so if you go to the journal nature is an entire sub domain on that website which talks about irreproducible research. And their concern is for all those papers trying to get published in nature, how do you guarantee that whatever insights or whatever results that somebody is trying to package in that paper. How confident are you that those are reproducible enough that they are worthy of being published and being published not just in any journal, but in nature.

So, they are so concerned that there is a sub site that they have created on the reproducibility and the lack of it. There are several causes for why this bias ultimately and I am just list I am going to list a few as we go along and not all of these have to do with omics, but I wanted to appreciate the broad idea and finally, towards the end we will talk about how to control for some of these, but the most important reason for why results are not a principle is something called a confirmation bias.

(Refer Slide Time: 09:44)



What are the causes of bias?

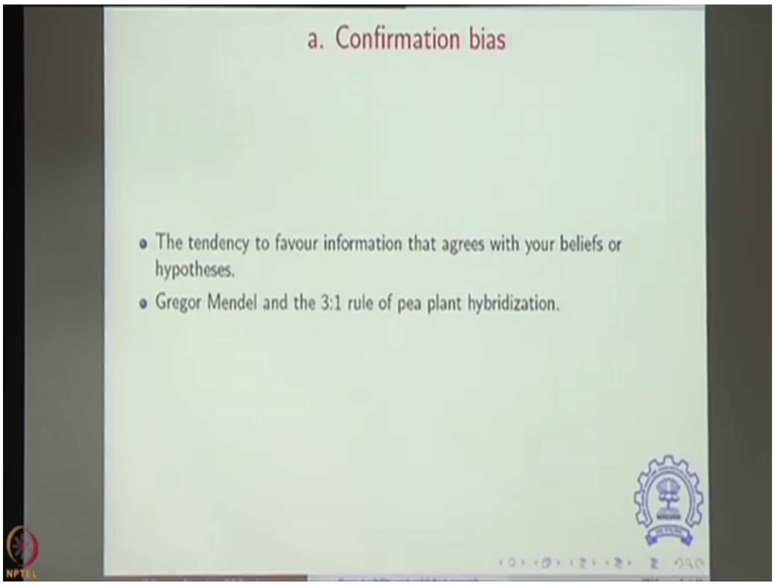
a. Confirmation bias.
b. Existence of design issues: Manipulation of the analysis.
c. Results are usually not independently tested.

Which is you already have a hypothesis going in your mind about what do you think is the underlying cause for this disease. And now when you are generating data you are only paying attention to those data points or that subset of data which you think confirms your hypothesis ok. And in doing so because you are only looking at a subset of data points you are probably missing out something more important, which might have told you something else about the diseased country ok. The other reason is very common which is that people manipulate that the analysis outright.

And a third situation is that the results are not independently tested you may have already had some discussion about this in terms of your omics pipelines, but a typical headache for example, is the tissue sample which you are processing for an omic study is not available to a different researcher. So, there is no way of validating whether your actual experimental workflow was executed correctly or not therefore, whether your targets are therefore, relevant or not.

So, let us look at these one by one. So, the oldest example of confirmation bias is actually somebody would be surprised with Gregor Mendel. To give you an idea of why you know; if Gregor Mendel were to publish in this day and age. In fact, he would not publish he would not be published because you would say that he cheated he pressurises data. Why? You sit and do that pea plant experiment which is famous for various expected to get a proportion of 3 is to 1 daughters with a certain phenotype.

(Refer Slide Time: 11:07)



a. Confirmation bias

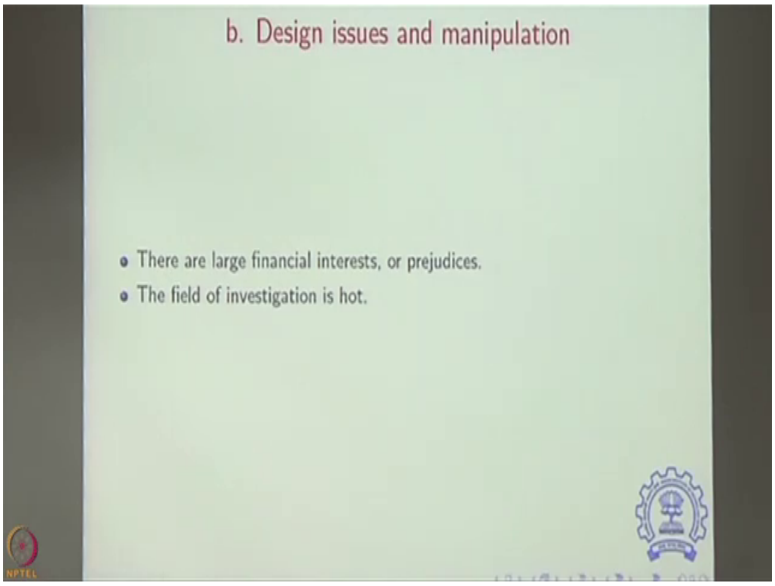
- The tendency to favour information that agrees with your beliefs or hypotheses.
- Gregor Mendel and the 3:1 rule of pea plant hybridization.

You do that experiment starting with pea plants your proportion of plants of daughter plants which have which are carrying a particular phenotype will not turn out to be 25 percent. 25 percent to give Gregor Mendel credit was something which you realized intuitively was some number to round off to that the 3 is to 1 proportion is probably a nice rounding off of numbers, but if you sat into the raw experiment yourself and you collected a whole bunch of plants and you categorize them by the length of the leaves and so on.

You would not get 25 percent of them having shot leaves versus long leaves for example. What does that tell you? It tells you that he was already biased towards reporting a result of 3 is to 1, he wanted a number 3 is to 1. Of course, there is no statistician there at that time to challenge him of course, these days there are when you send out an article for a review process.

So, when he says 3 is to 1 the whole bunch of people except 3 is to 1 as if it is the truth, but here is the thing that experiment could never have been reproduced ok. And he goes into this with a specific bias and the irony is the law of inheritance as he found it is an accurate description of how inheritance happens, but inherently he have a problem that this is not inheritable.

(Refer Slide Time: 12:30)



b. Design issues and manipulation

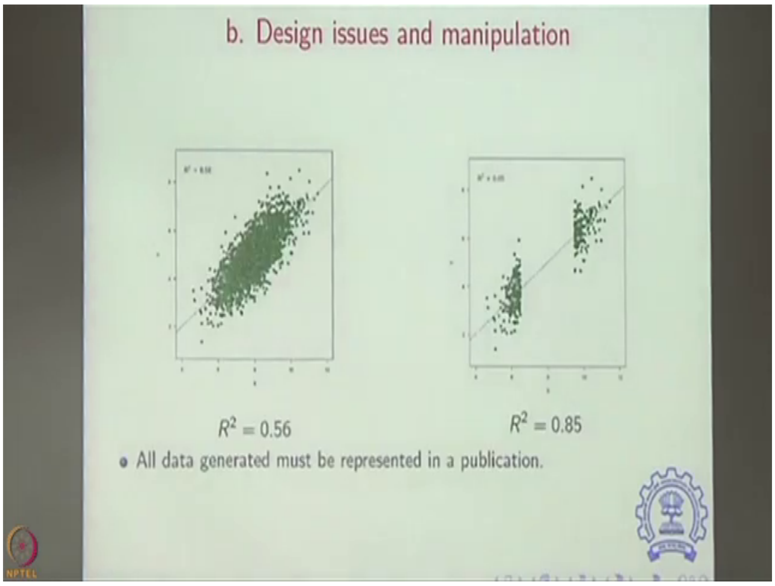
- There are large financial interests, or prejudices.
- The field of investigation is hot.

The second reason why most results are not reproducible is that there is manipulation of these data. And especially in large departments in the west, he was starting to see a paradigm where there is sponsorship of the research in the department by some large entity a pharmaceutical company.

And that starts influencing look at the whole process of which types of hypotheses will you look at, what data will you generate, what kinds of experiments will you actually perform and report and you are not free therefore, to actually carry out certain types of experiments because you are using the pharma companies money; you are not free to do what you want to do. And therefore, you are not truly reporting what you think is a relevant inside.

So, this is an example where manipulation happens and this is a spin off of this is at any given point of time one particular type of investigation suddenly becomes hot and everyone tries to do that and that is accepted as a standard protocol for data generation, but there is nobody challenging ok. Because there is nobody capable of challenging given that it is not a kind of facility that is available to you everywhere ok. What the truth is of that particular data set or whether there is some pressure is involved in analysis.

(Refer Slide Time: 13:31)

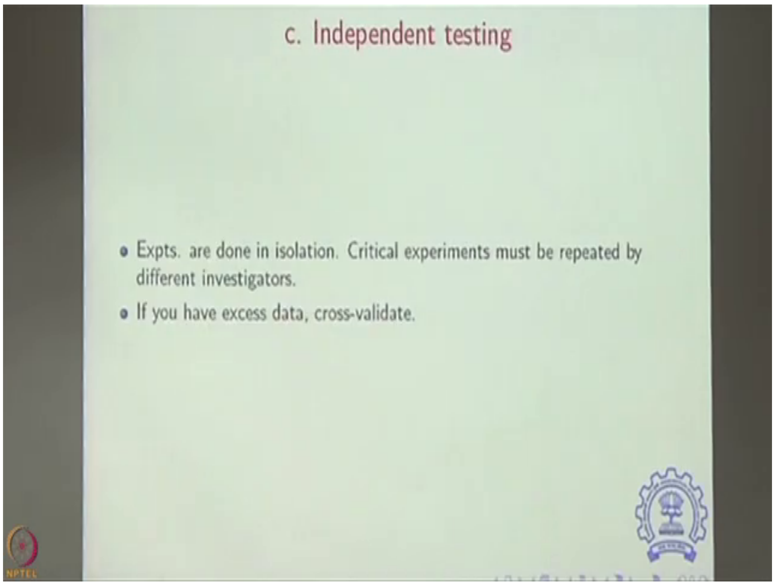


b. Design issues and manipulation

- All data generated must be represented in a publication.

And this is a simple example to point out what happens if you manipulate it outright by misreporting on in this case not reporting data. So, if you look at the plot on the right and you look at the reported as per value it is a simple straight line fit which I am reporting. I will swear that its a great line that as per value if you just go by a published as per value you think its a good model, but you see how you have arrived at a good model by just omitting a few data points, and the risk associated with selective reporting of data in a publication.

(Refer Slide Time: 14:03)



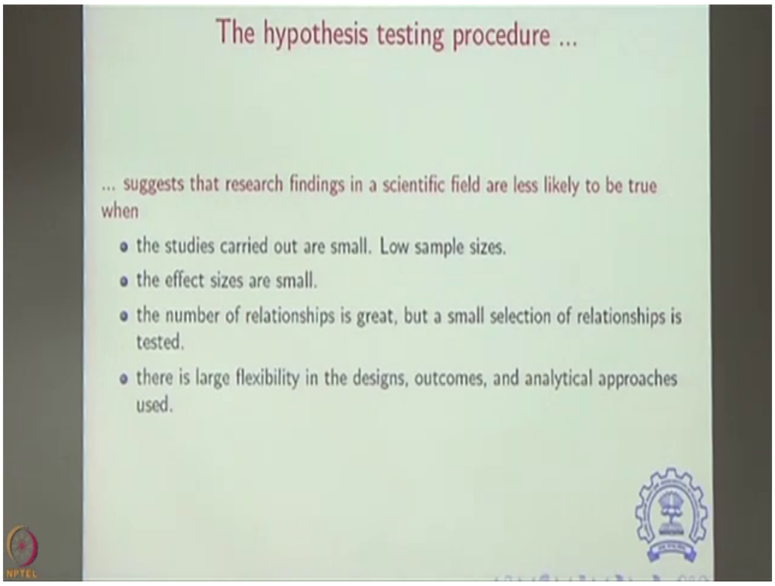
c. Independent testing

- Expts. are done in isolation. Critical experiments must be repeated by different investigators.
- If you have excess data, cross-validate.

And a final point on this line is that when you are talking about testing an independent testing, the fact remains that we are all working in isolation and therefore, there is no culture of materials being shared and being retested elsewhere.

As if you are saying that something is a critical hypothesis a critical gene for example, or a critical protein in some kind of a disease condition the first thing you have to do is for this protocol to be repeated by somebody else by different investigator who ideally does not have access to your raw materials. And therefore, is independently validating that your idea can be reproduced elsewhere.

(Refer Slide Time: 14:42)



The hypothesis testing procedure ...

... suggests that research findings in a scientific field are less likely to be true when

- the studies carried out are small. Low sample sizes.
- the effect sizes are small.
- the number of relationships is great, but a small selection of relationships is tested.
- there is large flexibility in the designs, outcomes, and analytical approaches used.

All of this comes down to how you set up for most part in especially an omics type of approach a hypothesis. The entire statistical framework of how we evaluate a data hinges around the fact that you ask whether somethings an important gene or not as a hypothesis and then you try to shoot down that process.

So, the easy way for you to remember this is actually how the legal system works. So, how does the legal system work whether it is in India or in the west what is it one tries to do. Somebody is what somebody is innocent until proven guilty and there is an additional clause there, what is that clause?

You are innocent until proven guilty beyond the shadow of a doubt that is critical beyond a shadow of a doubt you are innocent until proven guilty. So, the effort is on somebody to try to prove you guilty not to try to prove you innocent. So, somebody has to try and prove you guilty and that too that guilt has to be shown beyond a shadow of a doubt.

The shadow of a doubt is something that you would have heard of as a confidence level. So, I trust a result within a certain confidence interval and if I see a measurement beyond that particular range of confidence I say that this result that I have gotten is not consistent with the original hypothesis, I now go with an alternate hypothesis.

One simple example here is if I toss that coin 100 times and I get 15 out of 100 heads, 15 out of 100 as heads; then instead of calling the coin of fair coin I say my outcome was so extreme 15 out of 100, that I would rather go with the hypothesis that it is a biased coin and not go with the hypothesis that it is a fair coin I got an extreme result, beyond a confidence interval of sorts. So, therefore, the entire hypothesis testing procedure involves on taking one particular statement asking whether our observations are extreme relative to a range within which that statement can be true.

And then therefore, deciding on whether we are on the inside in which case that statement is or whether we on the outside in which cases what we have seen is something extreme. Now if you think about how you identify targets you look at gene one what is its fold change in expression is that fold change extreme. If that fold change is extreme then you say as you shortlisted and say that how extreme it was could not have been an event by chance. Therefore, you believe that it is an actually a good clinical target.

And for all the 1000s of genes that you look at you ask is that been some random variation in their expression levels within what range could that random variation have been. And you therefore, assign some confidence interval and when you see a measurement well beyond that ok.

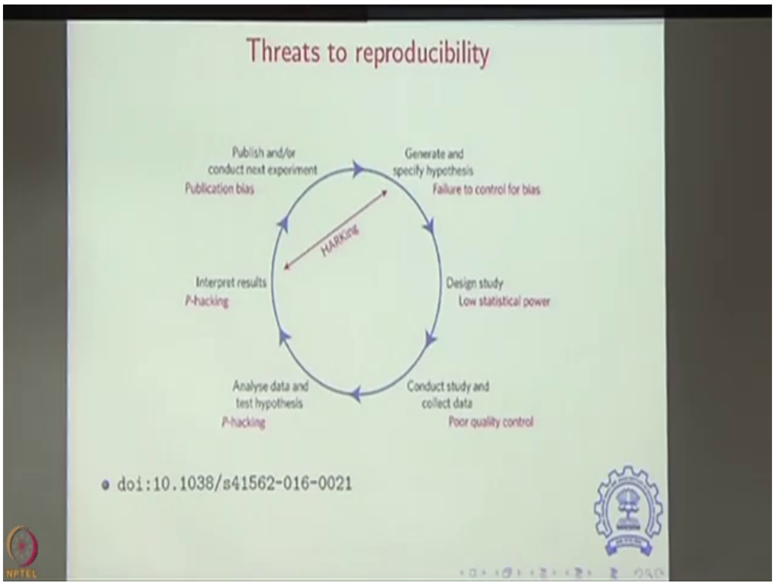
In terms of either fold up regulation or fold down regulation at this point you say; I saw something so extreme it could not have been by randomness could not have been by chance. Therefore, is something going on there as to why this particular candidate is over or under expressed ok.

However what we know is that these findings these kinds of studies in terms of statistics they are limited for several reasons. And in principle the reason that these findings are less likely to be true, is if whatever you are looking for is a small effect then odds are that you know you are going out on a limb by saying something is a significant gene when all you were out to do in the first place was to create a small effect odds are your phone around candidate.

Also if your effect sizes are small if your sample sizes are small and you have really no business saying that something is an extreme result. And if you are looking at many relationships if you in for example, in the omics case if you look at many genes and trying to make the argument that this gene is important that gene is important and I have got a whole bunch of genes to analyze one after another.

The larger that pool of things being tested odds are you making mistakes and I will come back to that in a minute, but take these points in the context of what threatens reproducibly I am not sure you can see this very clearly.

(Refer Slide Time: 18:40)



Threats to reproducibility

- doi:10.1038/s41562-016-0021

Some interesting buzzwords interesting terms which later on I really recommend all of you go look up ok. So, if you look at if you look at this cycle and in fact your cycle because its a publication cycle its how one does research with publications in mind.

So, if you look at the way you are supposed to operate you are supposed to operate at the top right there, where you generate and very specific hypothesis. You generate that hypothesis you are supposed to now design experiments collect samples ok. Design a study and when you design a study by the way you are supposed to design what is called a powerful study.

I will come to that power a little in a little bit, but your design of study which is a protocol which is going to help you discriminate between a control and a test case that is that is what you mean by designing a study well. And you then actually conduct the study and collect data so that is at the bottom. Again notice that write down I am putting down things which can go wrong at each level. So, when you are generating and specifying the hypothesis you might already be biased you are already looking out for a particular result ok

And you might already be biased in how you are therefore, trying to now carry out the study. If you think about how tobacco companies operated for a while in carrying out trials as to whether nicotine is addictive and this paper after papers saying from a tobacco company saying nicotine is not addictive there is a bias. Because they cannot be in a situation where they report that nicotine is addictive and harmful, because; obviously they will kill their own business.

So, you can see they are going to tweak the way their results are and the study is done true end up with a result which is; obviously, good for them. So, you generate and specify hypothesis you design a study and then you are worried about how capable you are of differentiating your control result from a test result.

You conduct the study and then collect data and usually a problem here is that you do not have good quality control in terms of your methodology. And then finally you get down to what is called data analysis and some of you if not all of you have probably heard of something called a p value how many of you have not heard of a p value ok.

We all heard of a p value. So, let us discuss that p value because a lot of omics critically hinges on the interpretation of a p value and what you can do with it ok. There is in fact, unfortunately a phenomenon called p hacking and again you really want to look up p hacking if you have not seen this before. Which comes about where people use the p value concept without having a true understanding of it and try to label some targets as being important and others is unimportant.

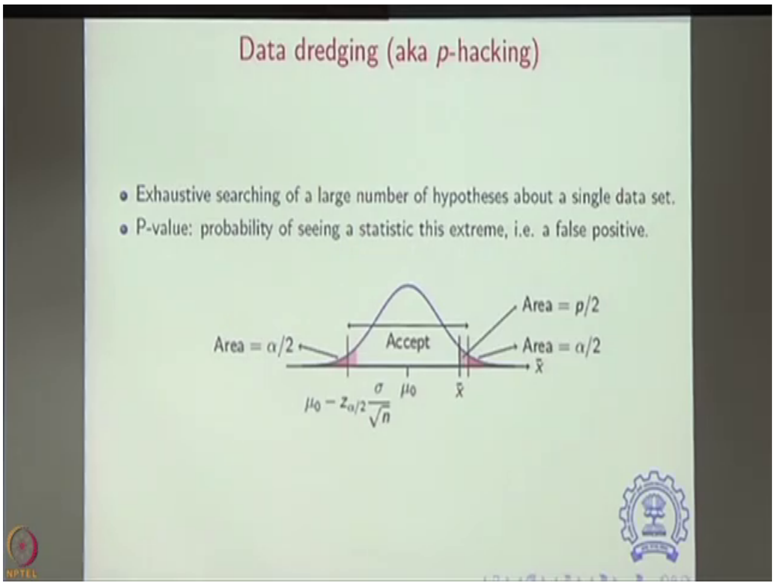
So, the interpretation of this data via hypothesis testing approach leads people usually once you get this data then you are going to do some follow up experiment which means you are in the next publication cycle, you have set up some follow up activity and now maybe you are testing these targets.

So, you are testing different tissue samples using this particular protocol you would normally get locked into some kind of a cycle where if you have already published 2 2 papers on a particular cell type and with a particular subset of genes. You are forced into publishing more just to offset the costing of the research that you are doing.

And this is where a bias can creep in because you are not in the position to say that what you have been doing all along this poor. You will notice that there is this double this is line as an arc cutting across ok. The circle and have written harking there harking hark refers to hypothesis after the results are known which is unfortunately what happens to most of us which is we generate the data and then we start asking what exactly are we trying to as an insight because at this point you have put in a lot of time and effort into your study you better come up with an insight.

So, therefore, you ask the hypothesis after the results are known which fundamentally is cheating. Because you know you can decide what you are going to call significant and what you want to call insignificant. And this goal post business if you go if you were ever going to have a goal post to decide what is a good target, what is a bad target; that should have been done before you even looked at data ok, otherwise you have the potential to cheat.

(Refer Slide Time: 22:39)



Data dredging (aka p-hacking)

- Exhaustive searching of a large number of hypotheses about a single data set.
- P-value: probability of seeing a statistic this extreme, i.e. a false positive.

So, this is p hacking business is the crux of most issues to do with reproducibility. So, let me quickly go through that concept ok; p hacking also has a nice phrase called data dredging. Basically involves you searching through a large number of hypothesis, typically with a single data set. So, the data set for example, could be that you have one genomic data set or one omics proteomics data set and you are asking the question which subset of candidates are important as targets.

How does our hypothesis testing procedure work? So, if you look at this what does this mean we expecting the fact that in an experimental condition this should have been the average value I should have seen if replicates wherever done. Now there is already a practical problem which is in an omics world they are probably not going to have the luxury of doing many replicates because it is a very expensive protocol to be executing and you may not even have tissue samples to play around with ok.

If you had the luxury of doing many replicates you would see a range of values for the same gene, that range of values would follow some kind of a distribution like this ok. And with your luck you may end up with the result anywhere under this distribution just like when you toss that coin 100 times, it is possible for you to get 50 heads out of 100, but its also possible for you to get 30 out of 100 and for that matter 70 out of 100.

These are all possible outcomes given that you are not in the business of doing the experiment again and again and again. Notice that statistics ideally says do things again and again and again and average them out because then you are more likely to find the true value.

Intuitively you know that that if you work with for example if I ask you what is the average height of a person; what is the average height of a human being. Either I can work with your one I can take one individual and use your height as a representative of the average, but you know that is not smart because I might end up with a short person or a tall person.

Instead the safer thing to do is what take an arithmetic mean of everybody out here and then use that as an estimate of the average you know that; you know that is a better representation of an average ok. So, the whole notion that results can occur under some distribution ok, but if this is what you expect then we know that between these 2 boundaries here, between these 2 boundaries we are in a situation where for example, if this were 50 heads out of 100 this could be 30 this could be 70.

Then we are in a situation where any outcome we see the one time I do my experiment between 30 and 70 any time I do the experiment once and I see an outcome between 30 and 70. What do I say them? What do I infer as the hypothesis result? I say that my result is close enough to 50, but I probably saw a result associated with a fair coin.

It is only if I see an extreme measurement sitting out here or sitting out here that I said look what I just saw is so extreme I saw 15 out of 100 or I saw 85 out of 100 that is so extreme it could not have been because of a fair coin..

Therefore, I should be in considering this coin to be a biased coin ok; that is how we interpret a typical hypothesis. Now if you look at this these 2 barriers which by the way of confidence intervals and for us typically what is a magnitude of this conference interval, what do we choose anybody has an idea what magnitudes do we choose you would have heard of a 95 percent confidence interval as a typical setting for confidence intervals.

So, these 95 percent of confidence what you are saying is 95 percent of the area under this curve probability which is probability is between this threshold on that, which means the 5 percent sitting outside. Now the headache is how do you interpret this 5 percent sitting outside. So, what does therefore, this dark red shaded area mean to you? That dark red shaded area is a situation where you could have had a fair coin, because the whole thing is associated with being a fair coin; these are the range of outcomes you would see with a fair coin.

You could have had a fair coin, but its sheer bad luck you see an extreme result like 15 out of 100 or 85 out of 100. And at this point you go with the other hypothesis that its a biased coin, but it could have been a rare outcome with a fair coin.

In which case as a technicality you have made a mistake, in your hypothesis why because you are gone with the other suggestion and the other hypothesis than with the original hypothesis you have gone with the assumption that the coin is biased then you should have gone with the assumption of the coin is fair and what you see is a rare event with a fair coin.

So, this 5 percent which is something we tend to take for granted is actually critical in a an omics test because it represents an error it represents an error in interpretation. So, this 5 percent is an error, where we should have gone with the null hypothesis as this is called we should have gone with that original hypothesis, but instead we are preferring to go with an alternate belief an alternate hypothesis.

So, 5 percent of the time we commit an error in an analysis. So, we commit mistakes 5 percent of the time so that is the take home message of that. But how many times did you carry out to study how many of how many hypotheses did you test in an omic study? How many genes are you testing? If you are testing 10000 genes you know study one after another and I just told you are committing a mistake 5 percent of the time then have you committed mistakes in your analysis of 10000 genes.

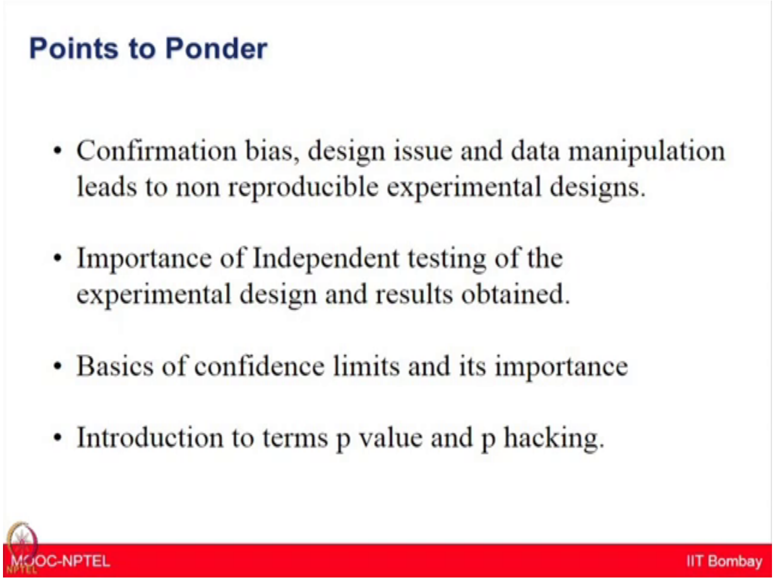
You would have somewhere in there you would have ended up with the wrong conclusion just by sheer bad luck. Because your error rate is 5 percent your error rate does not matter much if you are testing something once, but if you tested something 10000 times well you did not test one thing 10000 times, you carried out 10000 different tests one per gene. And if you tested 10000 things one after another you made a mistake by sheer bad luck at least 5 percent of the time which amounts to a large number of candidates potentially being wrong ok.

Therefore, if you were looking at your null hypothesis being this gene is not important and you are trying to shoot down the null hypothesis. So, if my hypothesis is this gene is not important because its expression fold change is one and I am looking at fold up or down regulation close to one, then any range around one corresponds to my null hypothesis being right. And what is the null hypothesis? That gene is not important and it is only when I see an extreme measurement would I now say this gene is important, but 5 percent of the time for a given gene I would make a mistake and I just did this analysis 10000 times one after another.

Which means most of the times I am actually calling something important as a gene candidate, which is up or down regulated it is not even statistically important to you. It could have occurred by sheer chance sheer chance why because we did the experiment few times and with sheer bad luck we are seeing extreme results. And we are being fooled into thinking that our insights are important when they are not ok.

So, the p value which is now so let us say our outcome is over here; if my outcome is over here you then ask the question how far away from this 95 percent threshold was I and how close am I to changing my mind ok. And this p value if this so by the way either 5 percent of the area outside this threshold on the right is slightly more than 5 percent of the area outside this threshold, that is intuitive because I just move the goalposts inside. So, there is more than 5 percent of an error that you committing the question is how close are you to this threshold what are the odds of this error being extreme.

(Refer Slide Time: 31:00)



Points to Ponder

- Confirmation bias, design issue and data manipulation leads to non reproducible experimental designs.
- Importance of Independent testing of the experimental design and results obtained.
- Basics of confidence limits and its importance
- Introduction to terms p value and p hacking.

MOOC-NPTEL IIT Bombay

So, I hope you are convinced that there are many minor, but very crucial considerations about a good experimental design. I hope you have learnt how to analyze your data and to make best sense out of it while thinking some very minute aspects of experimental design and data analysis. We also studied how conformational biases may lead to missing out some very important information from your data sets.

Also there are various reasons like confirmation bias, design issues, data manipulation along with the lack of independent testing which could lead to the non reproducible experimental design. And if you want to publish in very good scientific journals you need to ensure that all of these considerations have been met. I hope would you have also learned about the importance of confidence interval in the selection of a potential reliable candidate. We have also learnt about how p value could affect the interpretation of results on your data.

So, you need to be very careful about knowing some of the terminologies used in statistics how these tests could be performed. And how one should really make a meaningful interpretation from the this data set which is available to you, which is usually very large data set from the omics experiment, but what is the most significant leads out of that needs lot of consideration needs lot of considerable thought process and your understanding from experimental design to the various type of tests being performed.

And then only finally, you can come up with a reliable list of the proteins, biomarkers or Kennedy drug targets which could be meaningful for the future experiments. So, the next lecture would also be continued by Professor Noronha and he will further talk to you about various factors involved in good data analysis and experimental designs.

Thank you.