Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Dr. D. R. Mani
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay
Principal Computational Scientist, Broad Institute

Lecture – 20
Data analysis: Statistical Test

Welcome to MOOC course on Introduction to Proteogenomics. In the last two lecture, Dr. D. R. Mani provided you the basic information about how to now utilize your data generated to make it much more reliable and a statistically amenable for further analysis. So, by now you are comfortable in generating data using mass spectrometers, but even you have obtained the very complex datasets for variety of samples, it becomes very crucial that you are really working on the right data set for analysis.

You have done proper ways of normalization and now you need to employ the statistical tools for further analysis to find out the most significant hits out of a data. If your aim was to compare a disease verses healthy individual and we have run you know hundreds of patient sample with the individual healthy controls, now your next goal is to find out how many proteins are differentially expressed which are significant from one to other population.

In this slide, Dr. Mani's lecture, today we will focus on different types of a statistical test and it is applications in proteomics. So, as we all know that any data analysis requires the power of a statistics to strengthen the outcome. Some of the important and very common type of test which are employed for this purpose includes correlation, regression and different comparison strategies. So, correlation is a technique which can show whether and how strongly the pairs of variables are related whereas, Pearson correlation is important to test for the strength of the association between two continuous variables.

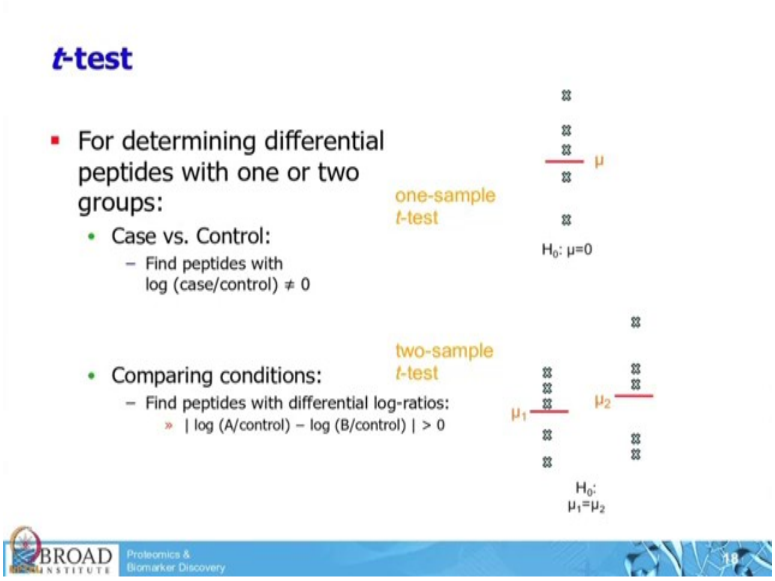
On other hand the Spearman correlation is a test for the association between two ordinal variables, they does not rely on the assumption of normal distribution. Then next comes the regression which can estimate if a change in one variable can predict the change in other variables and there are mainly two types of regression test. A simple regression which test that how the changes in predictor variable can predict the level of change in the outcome variable whereas, the multiple regression looks for the test how change in the combination of multiple predictors variable can predict the level of change in the outcome variable.

Another test which we often need in comparison of means which actually looks for the difference between the means of variable. Paired t-test it is another important one where it tests for the difference between two related variables. Independent t-test on the other hand looks for the differences between two independent variables. Next comes the ANOVA or the Analysis of Variance, this analyzes the difference among group means in the sample. But sometimes when data does not meet assumptions required for the parametric test, we have to then do nonparametric test which means the data is not required to fit a normal distribution.

A simple test Wilcoxon signed-rank. This test used to compare two related samples to assess whether their population mean ranks differ or not. Two sample test Wilcoxon Mann Whitney test this test is used to compare the outcomes between two independent groups whose data are not normally distributed. So, I am sure you appreciate that there a variety of a statistical test and options available to compare a data under which context, you need to employ which type of test becomes very crucial to understand and that is where today's Dr. Mani's lecture is going to give you more detail and examples that in which context you need to employ which type of a statistical test for a data analysis. Let us welcome Dr. D. R. Mani for his today's lecture.

This is a kind of a cartoon figure for how a t-test works.

(Refer Slide Time: 05:04)



t-test

- For determining differential peptides with one or two groups:
  - Case vs. Control:
    - Find peptides with log (case/control) ≠ 0
  - Comparing conditions:
    - Find peptides with differential log-ratios:
      - | log (A/control) – log (B/control) | > 0

one-sample
t-test
μ
H_0: μ=0
two-sample
t-test
$μ_1$
$μ_2$
H_0:
$μ_1$=$μ_2$

BROAD INSTITUTE Proteomics & Biomarker Discovery 18

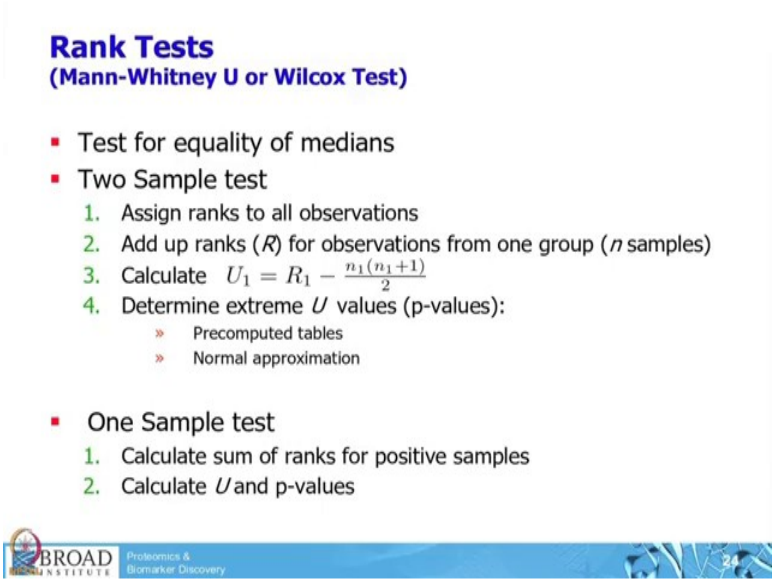
So, you have some mean and you have some distribution about the mean. So, based on the distribution about the mean you calculate a standard deviation and based on that you calculate a t statistic and if that statistic is extreme enough, you say then it is statistically significant. So, you use the extremeness of the statistic to calculate your p value. And so, the p value and the extreme value of the statistic are correlated. So, if you have a low p value then your statistic is more extreme than usual and so, based on that you can.

So, if μ was very different from 0 and if this distribution was not too big, then you would say this is statistically not 0. But if this was close to 0 or your variation was so, large that you could not tell whether it was 0 or not then you would not say it is statistically significant intuitively. Similarly for a two sample t-test you are looking a two groups and you want to see if the difference in means is large enough. So, it not it depends not only on the difference of the means, but also on how much variation you have around each mean. So, that is why you do not want to like just take the means and then see if they are different, you want to do a formal test to take the variation of those measurements around the mean into account. Is someone have a question?

So, regular t test usually ends up being very unreliable when you do it with a small number of samples like 5, 10; if you have that many samples then a regular t-test when you calculate the standard deviation depending on the actual samples you have measured, the standard deviation can vary quite a bit and when that happens the t statistic varies quite a bit and so, your p values can vary quite a bit and will not be robust, when you do a test with a small number of samples. So, there is a variant called a moderated t-test. So, what we here is, we look at all the data we have seen; we look at all the proteins we have measured and see what is the average variation for the proteins.

And then you pick a specific protein and if that protein is varying way too much, you say this look does not look like all the proteins I have seen in terms of variation. So, I am going to damp down the variation based on what I have seen in all my data. So, you kind of moderate your standard deviation calculations to make your test more robust. So, when you are working with small numbers of samples, it is a very good idea to apply a moderated test. When you have large numbers like 50 or 100 a moderated test under regular t-test would probably give you the same results.

(Refer Slide Time: 17:52)



Rank Tests
(Mann-Whitney U or Wilcox Test)

- Test for equality of medians
- Two Sample test
  1. Assign ranks to all observations
  2. Add up ranks (*R*) for observations from one group (*n* samples)
  3. Calculate $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$
  4. Determine extreme *U* values (p-values):
     - Precomputed tables
     - Normal approximation

- One Sample test
  1. Calculate sum of ranks for positive samples
  2. Calculate *U* and p-values

But if you have a smaller number of samples a moderated test is usually much more robust. So, the F-test like I mentioned is for multiple groups. So, you want to see whether any of these means are different from 0. So, there are a lot of nonparametric tests which I am going to skip for our reasons of time, but whatever you can do with a parametric test, you can also do with a nonparametric test. So, there are many so, the t-test whatever it does, it checks whether two means are the same or not. So, there is a the rank test that can do the same similarly for a two sample test and for an f-test and so forth. So, I will not go into the nonparametric tests, but either listed in the slides and you guys can take a look.

Student: Sir, these are both the one sample details and two sample details.

Yeah.

Student: Two sample means two groups.

Yes.

Student: Two comparing and one sample means.

Only one group.

Student: One group.

And you are checking whether the value of the mean.

Student: (Refer Time: 08:58).

Is same as 0 or not. So, that is where. So, the kind of data you have is very important. So, let us say you how ratios that come out of your mass spec processing software and you did not log transform it and you do a one sample t-test. You are always going to get things as everything will be significant because, 0 is not even a valid value you are ever going to get out of the software you are doing. If you just take ratios, the value is going to be greater than 0 all the way to infinity and when you do a statistical test to see whether it is more than 0 or not, it is going to be more than 0 by definition.

So, that date for a data that is only a ratio you would not do a one sample t-test, because it does not make sense conceptually. But once you log transform it then the question you are asking is, whether it is up or down regulated. Because if it is up regulated, it is greater than 0 whether it is if it is down regulated, it is less than 0 and if it is not regulated it is around 0. So, that is the reason why we kind of do a log transform because it makes things symmetric and may makes many of these tests applicable. So, if you took only a ratio, you cannot apply a one sample t-test, but if you took the log of the ratio then you can.

Karl: Any test needs replicates.

That is true. I am assuming you will have some. So, Karl is mentioning that I should point out that to do any tests you need replicates because why do you think you need replicates? What are we calculating?

Student: Variants

Variants exactly we are calculating standard deviation and you cannot calculate standard deviation with a single number, you need at least two. So, in to do any of these statistical tests, you have to have replicates without replicates you cannot do either a one sample or a two sample or any of the tests. So, if you measured your sample only once or there is a group for which you have only one sample. So, you are doing breast cancer analysis and your HER2 positive group, you have only one sample and you cannot say anything about that group because you do not have enough replicates to do a formal test.

So, any number you see, you can tell whether it was 0 or not. So, in the worst case; if all you have is one sample and you desperately want to say something about it what you can do is, you can plot the log ratios and then pick the most extreme you assume it is a normal distribution and in a normal distribution anything that is beyond two or three standard deviations is considered an outlier or not is a is an extreme value. So, you can do something like that and still come up with a list of proteins or genes that are different if you have only one replicate, but that should really be a an extreme situation that you hopefully never run into, but it happen sometimes. So, in for all these tests, you need replicates and the more you have the more robust your results will be.

Student: If we are doing cell line work and studying the effect of a certain drug in cell line

Yeah.

Student: So we have one control and one drug treated sample.

Yeah.

Student: And the sample size is also very less and in one two groups are there. In this scenario what should we know.

So, the question was they have a drug treatment experiment where the treated and untreated are will have few samples like very small number of samples and they want to find what is that what is different. So, you would do a moderated test assuming you have at least two for both treated and untreated and you would do a moderated t-test. So, moderated tests work well with small numbers of samples, to is probably cutting it close, but if that is what you have that is what we have. So, in our group there are lot of studies where we have only two replicates and we work we have done moderated tests with those.

There are a lot of other tricks you can use when you are working with just two replicates, you can try to remove things that are not reproducibly measured in both the replicates. You can kind of try to trim your data to do as little as few tests as possible, but you can still do a moderated test.

Student: Can we sample number of the various sizes, can be compared in a single group?

What do you mean?

Student: Like HER2 would be 3 other type 5 or 7.

So, different number of samples per group.

Student: Per group.

Yeah. So, that is one of the beauty of the test is that you put all the samples of a group together and calculate mean and variance. So, if you had a group with 50 samples, the mean and variance calculation will be very robust; you will be fine. If you have a group with only two samples the mean and robust variance can still be calculated, but it will not be as robust. So, if you compare a group two groups where you had larger number of samples and the marker was different, you will get a lower p value. But if one of the groups has fewer samples, you can still do the comparison, but your p value will be higher because the are the way the how the definite you are about the estimation of your parameters is more uncertain with fewer samples. So, you can do it, but you will get correspondingly higher p values.

Student: Sir, as you said that the extreme values are actually outliers, but extreme values, but

Yeah.

Student: Could there be situations where these extreme value would actually true and should not be considered as a outliers and because?

Well outlier was probably not the right word to use I think. So, when you have a distribution of things, things that are at the extremes are considered to be unlikely in most situations. So, that is why they are considered to be to validate the alternate hypothesis.

Student: Ok.

So, in that way they are outliers, but they are still part of the distribution. So, if you have only one measurement, it is quite possible that something that was an outlier, if you measured it again would be in the middle because it was a very low abundant protein and you can' measure it well and you ended up measuring it having one poor measurement. So, that is the problem with using only one sample for analysis and that is the reason why you want more samples. So, if you have two replicates and one was an extreme, the other was right in the middle, then the statistical test is going to say wait. It was over there and this is here; that means, I cannot really conclude that this is a extreme value or a central value. So, I am going to say this is not different based on the variance.

But if both measurements came out in the outlier, then you have more evidence that this is actually an extreme value and so, it would be statistically significant. So, that is where the variation of your values is taken into account in calculating your statistical significance. If you have only one replicate to bag, you can do it.

Student: If we have only one replicate, but if it still I was asked to draw a model around how it could be behave.

That is what I am saying the model I am saying is a normal distribution.

Student: But it may not be normal.

Yeah, you could assume other distributions by looking at it, but again you have to fit a distribution to a observed set of values and then use that distribution to calculate which ones are extreme.

Student: Like that was the for the genomics what kind of models may work?

So, most likely people use normal sometimes they use a log normal distribution, which is basically if you had direct ratios, you could use a log normal or a chi square distribution and I think he mentioned a Pareto distribution. So, if you have sort of normally distributed, but with very heavy tails, you can use the Pareto distribution. The problem is the more complex or exotic the distribution becomes the more things, you have to estimate in order to fit the distribution. So, for a normal distribution you just need mean and standard deviation and the more parameters you have to estimate the more uncertain your fitting becomes and so, the more kind of questionable your results become yeah.

Student: Sir, replicates you said earlier talking about technical or biological.

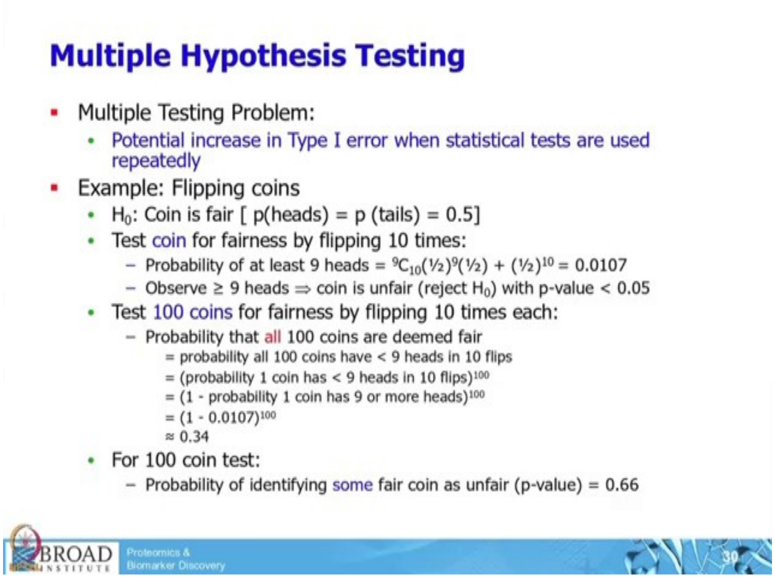
When I mentioned replicates, I meant biological replicates if you have technical replicates, then technical replicates are correlated because it is the same thing you are running again and again. So, suppose you had a study where you had 5 biological samples and you ran 3 replicates of each.

Student: hm.

You cannot put the 15 together and do a t test. Because the t test assumes that your samples are independent, here they are not because three of them came from the same biological sample. So, there you have to do what is called correlated analysis you have to account for the correlation. So, there are a couple of models you can use for that they are more complicated and I will not go into it today. But I the limma package I mentioned earlier, you can specify that these three are technical replicates of that sample and then you can take that into account in your modeling. So, it all boils down to you have to calculate variance and the question is the variance the same for all the samples or our groups of samples how different variances?

So, if there is a variance matrix you calculate and the structure of the variance matrix is what decides how you deal with replicates and technical replicates and biological replicates. And so, when you have technical replicates you have to specify that to calculate the appropriate structure for the variance matrix. So, there are models called linear mixed effect models, that are very versatile, they can account for things like correlation because of replication, they can account for time series, they can even take missing values into account and they provide some kind of moderation of how you calculate variances and things like that. But they are harder to set up and many times they do not converge and you have to like fiddle with the models to make sure it works and it is not straightforward for a for a person who has not worked with them to just like use it easily, which is why I do not want to cover it now.

(Refer Slide Time: 19:37)



Multiple Hypothesis Testing

- Multiple Testing Problem:
  - Potential increase in Type I error when statistical tests are used repeatedly
- Example: Flipping coins
  - H_0: Coin is fair [p(heads) = p (tails) = 0.5]
  - Test coin for fairness by flipping 10 times:
    - Probability of at least 9 heads = $^{9}C_{10}(½)^{9}(½) + (½)^{10} = 0.0107$
    - Observe ≥ 9 heads ⇒ coin is unfair (reject H_0) with p-value < 0.05
  - Test 100 coins for fairness by flipping 10 times each:
    - Probability that all 100 coins are deemed fair
      = probability all 100 coins have < 9 heads in 10 flips
      = $(\text{probability 1 coin has} < 9 \text{ heads in 10 flips})^{100}$
      = $(1 - \text{probability 1 coin has 9 or more heads})^{100}$
      = $(1 - 0.0107)^{100}$
      ≈ 0.34
  - For 100 coin test:
    - Probability of identifying some fair coin as unfair (p-value) = 0.66

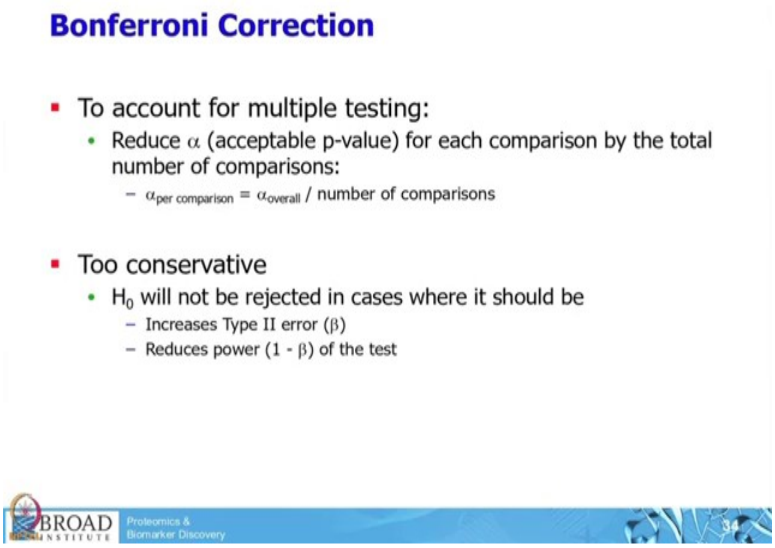
So, this is one thing that I want to mention and after that I think I should be done. So, we have mentioned that we are going to do a statistical test, to figure out whether a marker is different in cancers versus controls, let us say. So, you have you just measured 10,000 proteins. So, you want to find whether each protein is a marker or not. So, you do the test 10,000 times and you say anything that has a p value of less than 0.05 is a marker is statistically significant and is a marker. So, when you repeatedly apply the same test a large number of times, just by sheer chance; you will get things that are that pass the test.

So, you have a 5 percent chance that something that was not a marker, will have a p value less than 0.5, So, an alpha of 0.05 means that you are false positive rate or false positive probability is 5 percent. So, just by sheer chance by repeatedly doing the test on different markers, some number of them are going to be marked as false positive even though are marked as marker even though they are not. This actually gets significantly worse if you look at the example here. So, an example I have here is you want to decide whether a coin is fair or not.

So, you flip the coin 10 times it is a 10 yeah. So, you flip the coin 10 times and if you get nine heads, the probability of that happening is 0.01 if you do the calculations using the binomial theorem. So, this is definitely less than 0.5. So, if you get 9 heads, then you can conclude that your coin is biased, it is giving more heads than tails. So, if that is you how you set up your hypothesis test and you test 100 coins, there is a 65 percent chance that you will find at least one coin that is biased even though all your coins are fine. So, if you repeat the test many times, you will very likely find things that are different just by random chance.

To avoid that what we do is when we do a lot of tests like in genomics and proteomics we do this thing called multiple testing correction. We want to correct for the fact that our p values do not quite represent the real false positive rate.

(Refer Slide Time: 22:14)



Bonferroni Correction

- To account for multiple testing:
  - Reduce α (acceptable p-value) for each comparison by the total number of comparisons:
    - $\alpha_{\text{per comparison}} = \alpha_{\text{overall}}$ / number of comparisons
- Too conservative
  - H_0 will not be rejected in cases where it should be
    - Increases Type II error (β)
    - Reduces power ($1 - \beta$) of the test

BROAD INSTITUTE
Proteomics & Biomarker Discovery
34

So, there are lot of ways to correct for the multiple testing, I will just mention two. So, one is called the Bonferroni correction, this is the most conservative correction. If you do this correction and your marker still is significant at 0.05, then no reviewer will come back and question you. What you do here is suppose you did 100 tests, then you take your p value that you get and multiply it by 100. So, this is assuming that each of your tests are independent and you basically correct by multiplying your p value by the number of tests that you do.

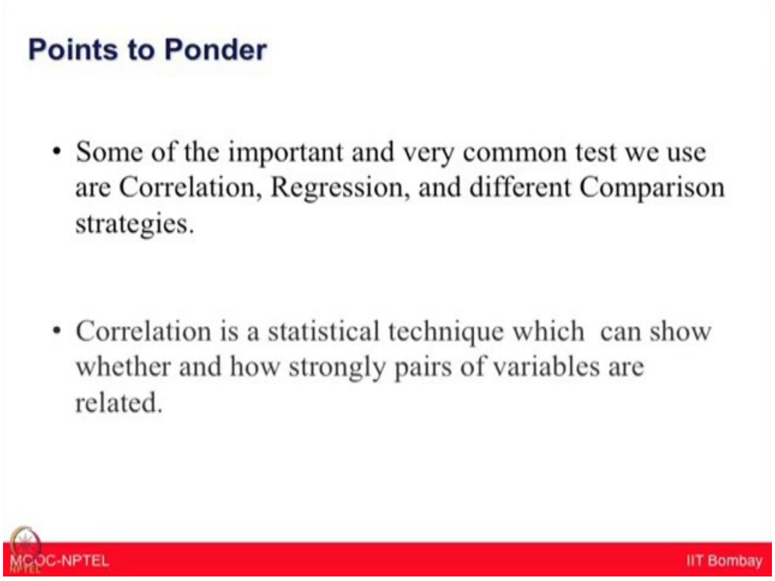
So, you can see that if you have unadjusted p value of 0.05 and you do a 100 tests, then multiplying by 100 is going to make the p value basically go beyond 1 which means it is not significant. So, this is the most stringent correction and many times this is too conservative especially with noisy data in genomics and proteomics, if you apply this test you will basically never ever get any markers that are significant. But the reason that happens, it happens is because this assumes that all your tests are independent that they are not related to each other, but we know that genes and proteins have relations.

So, now if you are they are part of the same pathway, they kind of behave similarly and there are lot of correlations and relations between genes and pathways that you have measured. And so, this in fact, does not quite hold. So, why do you want to be so, conservative when you know that your testing was not exactly like the assumptions that are made here? So, more relaxed correction is called the Benjamini Hochberg F correction. Here what you do is you define this thing correction factor that you apply to all your p values. So, intuitively what you do is you sort your p values from smallest to largest; for the smallest one, the lowest p value you got you do the Bonferroni correction.

So, if you did a 100 tests, the lowest p value you multiply by 100. The next one you multiply by 99, the third one you multiply by 98. So, you keep relaxing the stringency which you correct as you go down the sorted lists of p values. So, intuitively that is what happens and the paper this is actually almost universally used nowadays and the paper shows that if you did that and you set a p value threshold of 0.05, then your false discovery rate in other words how many things that are false in your entire list when you draw the cutoff there is 5 percent or whatever value you pick.

And so, this is a reasonable one to use and many times in small experiments where he mentioned, you have a drug response experiment of 2 versus 2 even the FDA Benjamini Hochberg correction may result in p values that are I do not know 0.1 or 0.2 and in those cases we just use the p value as a ranking mechanism.

(Refer Slide Time: 25:25)

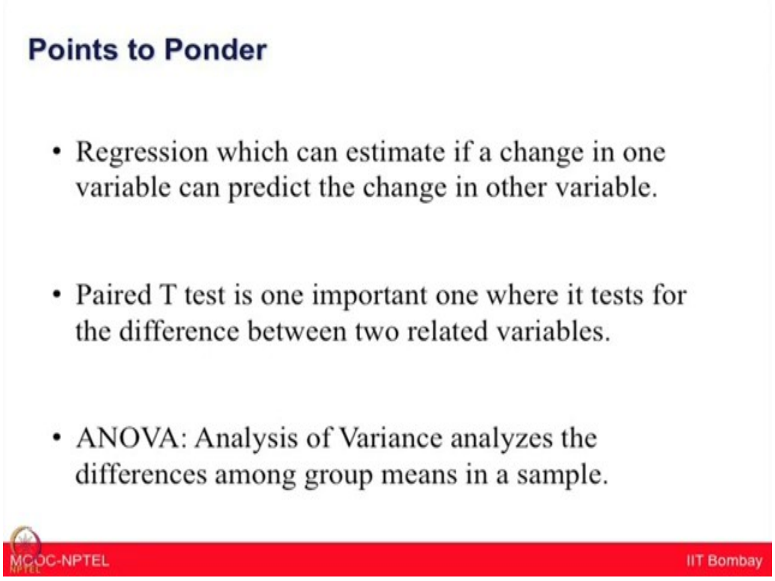


Points to Ponder

- Some of the important and very common test we use are Correlation, Regression, and different Comparison strategies.
- Correlation is a statistical technique which can show whether and how strongly pairs of variables are related.

MOOC-NPTEL IIT Bombay

(Refer Slide Time: 25:38)



Points to Ponder

- Regression which can estimate if a change in one variable can predict the change in other variable.
- Paired T test is one important one where it tests for the difference between two related variables.
- ANOVA: Analysis of Variance analyzes the differences among group means in a sample.

MOOC-NPTEL IIT Bombay

In conclusion, you have now learnt how to choose a right type of a statistical test and how to correct test can give the significant outcome from a data set. You also learnt increasing the number of samples, provides you much more increased confidence and the statistical power to your data. We will continue more interaction with Dr. Mani and in the next lecture he will talk to you about machine learning and clustering.

Thank you.