

An Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Dr. D. R. Mani
Department of Biosciences and Bioengineering
Principal Computational Scientist
Indian Institute of Technology, Bombay

Lecture – 22
Hypothesis testing

Welcome to MOOC course on Introduction to Proteogenomics. In today's lecture, Dr. Mani will be going to talk about hypothesis testing. Hypothesis testing is a statistical method that is used in making a statistical decision using experimental data. Hypothesis testing is basically an assumption that we make about the population parameters. Today's lecture, we will also explain you about null hypothesis and alternative hypothesis and then Dr. Mani will talk about the P-value and the power of a test. Further he will give you a brief idea about false negative test, the type I error, when we reject the null hypothesis although that hypothesis was true, what is the false positive test, which is the type II error, when we accept the null hypothesis, but it is false.

So, there are many considerations for hypothesis testing which I think is very crucial to learn as you are going along in generating big omics data sets. It is important also to understand the various statistical tests and ways to look at your data analysis much more carefully. Let us welcome Dr. Mani for today's lecture.

The next topic I wanted to address was hypothesis testing. So, this addresses questions about I have a marker that has some observed values in my cancer samples and some other values in my normal samples, is it a differential marker. So, questions like that fall into hypothesis testing, I will give a brief kind of formal overview and then give you some examples and what kind of tests you can use. There are a lot of slides. I am not going to go over most of them, I will just try to restrict it to like a small subset, so we can have more time for discussion and kind of trying to give a more practical overview.

So, hypothesis testing is use of evidence to come to a conclusion in some way.

(Refer Slide Time: 02:36).

Hypothesis Testing

Guilty vs. Innocent

In a court, a defendant is innocent (**null hypothesis**) until proven guilty. Strong evidence "beyond reasonable doubt" is required to convict the defendant (accept **alternate hypothesis** by **rejecting null hypothesis**).



 **BROAD INSTITUTE** Proteomics & Biomarker Discovery

So, I usually use some you know reasonable real world example, one kind of thing you can look at is, you are in a court of law and you want to decide whether somebody is guilty or not. So, in general the defendant is innocent until proven guilty. And so innocence is considered a null hypothesis. In other words; that is the standard hypothesis that is what is usually expected to be true in the world. And if you have a strong evidence beyond reasonable doubt that the defendant is not innocent, then you kind of accept the alternate hypothesis that they are guilty.

So, in cancer setting null hypothesis would mean that the gene or protein you are looking at, is not a marker and alternate hypothesis would be that your gene or protein is a marker for the distinction you want to make. So, there is the hypothesis you are making and then there is actual truth in the world that you are in.

(Refer Slide Time: 03:44).

The Hypothesis & Truth

Null Hypothesis H_0 , Alternative Hypothesis H_a

- Null Hypothesis usually represents status quo, no change, no effect
- Alternative hypothesis represents a change from the previous condition
- Null and alternative hypotheses are mutually exclusive
 - H_0 : Person is innocent.
 - H_a : Person is guilty.
- States of nature ("Truth") represents the true nature of things.
 - Actually innocent (H_0 true - H_a false)
 - Actually guilty (H_0 false - H_a true)



So, the hypothesis are, the null hypothesis is always represented as H_0 that the person is innocent, that is the null hypothesis that the person is guilty is the alternate hypothesis. But then in reality the person may be actually innocent or actually guilty. So, we do not know that; we do not know the state of reality, we are trying to do a test to kind of deduce what the state in reality would be based on some evidence we have seen, we do not know the truth.

(Refer Slide Time: 04:14).

Decisions: Testing the Hypothesis

- Collect and analyze data to support or refute the hypothesis:
 - Reject $H_0 \Rightarrow$ Sufficient evidence to say person is guilty
 - Fail to Reject $H_0 \Rightarrow$ Insufficient evidence to say person is guilty

Decision	State of Nature	
	H_0 True	H_0 False
Reject H_0	Person is innocent Sufficient evidence of guilt Jail Innocent person	Person is guilty Sufficient evidence of guilt Jail Guilty person
Fail to reject H_0	Person is innocent Insufficient evidence of guilt Release Innocent person	Person is guilty Insufficient evidence of guilt Release Guilty person

And so it results in a 2 by 2 table basically. So, the state in of nature or the actual truth is that H_0 is true. In other words, one possibility is that the person is really innocent and the other possibility is that the person is not innocent. So, those are the two situations that can happen in reality; that is the truth. And when you make your; so, you look at some evidence you do some test, you do some calculation or whatever and you come up with a decision to say you reject H_0 . In other words, you say the person is not innocent or you fail to reject H_0 . In other words, you do not have evidence to say that the person is not innocent.

So, those are you the decisions you can make based on the analysis you have done. And so you have a 2 by 2 table which, so suppose you can reject H_0 . In other words, you think you have enough evidence to say that the person is not innocent; if you did that and the person was actually innocent, then you are going to jail on innocent person. So, this is an error. Similarly here if the person was not innocent and you said the person is not innocent you jailed a guilty person and that is fine; that is correct.

So, similarly here for the bottom row the options are; you release an innocent person which is the correct thing to do or you release a guilty person which is not the right thing to do. So, these two diagonal cells are errors. So, this is a false positive, this is a false negative; the other two are correct.

(Refer Slide Time: 05:58).

Hypothesis Testing

1. Set up null (H_0) and alternate hypotheses (H_A)
 - Null Hypothesis represents status quo, no change, no effect
 - Alternative hypothesis represents a change of interest
2. Define a test statistic $T(X)$ and a Rejection Region (RR):
 - RR represents unlikely outcomes when H_0 is true
3. For given data X , reject H_0 if $T(X) \in RR$

Notes:

- Hypothesis is always stated using **population** parameters
- Test statistic T involves parameters computed **only** from the data
- A "successful" test results in rejecting the Null Hypothesis
 - » Null hypothesis is never "accepted"

So, if we look at the full table, this is called a type I error, this is called a type II error in statistics and you derive your P-value based on your type I error. So, the type I error gives you the probability that you reject H_0 . In other words, the probability that you say the person is guilty given that the person is innocent.

So, in other words, the probability of a false positive is represented by alpha and that is called the P-value of a test. So, the test is whatever you are doing to come up with the conclusion whether the person is guilty or not. Or in your marker case, whether a protein is a marker or not and the P-value is given by that probability, and the probability of a type II error is when you fail to reject H_0 . In other words, you have a marker that is really a marker, but your test is not strong enough to find the marker. So, that is the probability β and the power of a test is how well can you find those markers. If there is a marker, what is the probability that you will find the marker? So that is like 1 minus the error rate is called the power of a statistical test.

So, in general when you do these statistical tests, you want alpha to be as small as possible. In other words you do not want false positives, but you want the power to be as large as possible. In other words if there is a marker you want to be able to find it, so that is kind of the goal of how you do a test and the power of a test especially depends on how many samples you have. So, if you are using a test with a specific value of alpha what probability you have of finding a marker that is a real marker depends on how many samples you have. And so many times when you look at study design or when you are designing a study with patients and stuff like that, you will need to know how many patients you want and then that depends on what power you want to achieve in your study.

So, do you want to find all 80 percent of markers that have at least two fold difference and then based on some calculations you can come up with how many samples do I need to achieve that. So, that is where the power is most commonly used, but in a discovery like setting where the number of samples is fixed by extraneous forces; like how much money you have to spend on the project, how many samples you have access to, how much how many post docs you have to run your samples and things like that then the number is usually fixed by other factors. In that case, you just look at the power, the P-value of a test to figure out how good your markers are and there you kind of just take what you get. You cannot say I want all markers with or some percentage of markers

with some characteristics, that is not possible in a discovery study which most of the CPTAC studies or discovery studies, so I mentioned P-values. So, actually might be one more.

So, in hypothesis testing what you do need to do is, you need to first define a null hypothesis and an alternate hypothesis. So, if you are doing finding differential markers, the null hypothesis is that the marker is not different between the groups you are looking at, and the alternate hypothesis is the marker is different. This is important in many situations, but for marker analysis maybe not. But in many other situations you have to be clear on what your null hypothesis is, and what your alternate hypothesis is before you go ahead and do a lot of the analysis. Otherwise you tend to tune your null hypothesis to your convenience which is not a good thing to do.

Once you have defined that you define a test statistic. So, that is based on the test you pick. If you take pick the T test, there is the T statistic; if you pick the rank test, there is the W statistic. So, there is a statistic which is a number you calculate using the data. So, the high null and alternate hypothesis are defined using population parameters, its uses the entire universe you are looking at to define the parameters. So, you can say this is a marker in my population having cancer and no cancer, not just the samples you are looking at.

So, that is the way you define your hypothesis. But once you have that you calculate a test statistic only using the data you have, because that is all you have to make your decision. So, the test statistic is data specific and then, you have some kind of a rejection criteria. So, you say if my P-value is less than 0.05, I am going to reject the null hypothesis and say this is a marker. So, for some protein your P-value is 0.003, then that is below the threshold that you set, and so because of that it is a marker. So, you need a specific alternate and null hypothesis clearly defined that the test you pick will specify the test statistic and you need to pick a rejection region which is the P-value cutoff that you need and in ideal world, all these would be done before you do any analysis.

(Refer Slide Time: 11:32).

P-values

P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP, REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥0.1	

skcd.com



And so when you run a test, you get a P-value and this is like a cartoon that kind of says how people generally interpret P-values. So, P-value is the probability of a false positive in a situation where you are looking at lots and lots of markers with lots of noisy data, with not too many samples per protein you are looking at. Many times the P-value is just a ranking mechanism, so it ranks your markers in order of importance on how good a marker they are for whatever group you are trying to separate.

So, many times people tend to assign more significance to a P-value and that can result in your, may the threshold of 0.05 is many times considered sacred. So, what if it was 0.49 or was 0.51, is it significant or not? And so we generally tend to use P-values as a ranking mechanism, its fine if it is less than 0.05 or if it is very small then it is good. But many situations if you have small number of samples and noisy data many of your P-values will not be statistically significant and these are, I will also talk about multiple testing corrections. So, when you do multiple testing correction, you have to adjust your P-values.

So, if many times your regular P-values will be significant in other words less than 0.05, but your adjusted P-values will not be. So, in those cases, are you going to say you do have no markers or what do you do? So, in those situations we generally tend to just use the adjusted P-value as a ranking mechanism. I will get into it again later, but I thought we, was mentioning here, I can skip these.

(Refer Slide Time: 13:34).

Characterizing Hypothesis Tests (Classifiers)

Decision	State of Nature		
	H ₀ True (-)	H ₀ False (+)	
Reject H ₀ (+)	False Positive (FP) <small>Type I Error (α)</small> <small>Arrest Innocent person</small> <small>Jail Guilty person</small>	True Positive (TP) <small>Arrest Guilty person</small> <small>Jail Guilty person</small>	⇒ Positive Predictive Value $PPV = TP / (TP+FP)$
Fail to reject H ₀ (-)	True Negative (TN) <small>Release Innocent person</small> <small>Release Innocent person</small>	False Negative (FN) <small>Type II Error (β)</small> <small>Release Guilty person</small> <small>Release Guilty person</small>	⇒ Negative Predictive Value $NPV = TN / (TN+FN)$

↓↓ ↓↓
Specificity **Sensitivity**
 $TN / (FP+TN)$ $TP / (TP+FN)$

- Result is positive (+) when H₀ is false (Nature) or H₀ is rejected (decision).
- Other relations:
 - False positive rate (α) = 1 - specificity
 - False negative rate (β) = 1 - sensitivity
 - Precision = Positive predictive value
 - Recall = Sensitivity



So, this slide summarizes the types of errors and basically gives names for each of the items in. So, like I said this is a false positive, this is a false negative, it is a true positive true negative and using these numbers you can calculate many different characteristics of your test. So, you might have heard things like sensitivity and specificity of a test or the positive predictive value or negative predictive value of a test. They are all defined based on this table.

So, when you have actual samples, each of these cells will have numbers in them. So, how many false positives did you have? I had 5, how many true positives? 15. How many true negatives? Some number. So, based on the number of samples you have each of these cells will be filled with some counts and using those counts, you can calculate all these various characteristics of your test and many times you will encounter things like that and people will say oh my test is very sensitive, but not specific and vice versa and in some situations, it may be good; in some situations, it may not be good. It depends on how you, what your question is and what you are analyzing, but this is to kind of show that how to calculate these numbers, and I think if you actually go to Wikipedia, there is a bigger version of this table with way more values included. There are all kinds of variants that people calculate which you can take a look at. So, just an example for false positives and false negatives; I guess nobody.

(Refer Slide Time: 15:28).

Parametric & Non-parametric Tests

- Parametric: Theoretically derived sampling distribution
 - t-test (mean)
 - Chi-square test (frequency)

- Non-parametric: Empirically derived sampling distribution
 - Sampling distribution generated by
 - Combinatorial analysis
 - Random permutations
 - Mann-Whitney test (median)
 - Kruskal-Wallis test
 - Kolmogorov-Smirnoff test (distribution, goodness of fit)



So, once you have the concept of hypothesis testing and you want to find markers then the question arises; so, what is my test, in other words, how do I calculate my test statistic and that depends on which test you use. So, there are many tests you can use and they fall into two major categories; one is called a parametric test. So, here you assume that the data follows some kind of a statistical distribution like a normal distribution or some other t-distribution, some statistically defined distribution and you use the characteristics of that distribution to come up with the statistic and the rejection threshold.

So, because of that even if you have fewer number of samples, because you are assuming the distribution you can calculate these numbers relatively robustly and you can do the test, but many times you may not want to make the assumption that your data is normally distributed or has a t distribution. In those cases, you would end up using a nonparametric test. So, in a nonparametric test you calculate a number based on the data you have seen; like you rank your data and see how many are below a specific rank and above a specific rank, so you can compare ranks. So, when you do things like that you have a nonparametric test. But if you have small numbers of samples generally a nonparametric test has less statistical power, because you have to calculate more things.

So, you end up getting less statistical power and the variation in the data is cannot be like easily taken into a kind of account by the, then the statistical distributions that you could

have used if you did a parametric test. So, both have their applications and you use them in different settings. When you have relatively larger number of samples like more than 25 or 30, you can use the parametric test. If you have small numbers of samples, very small numbers of samples then a nonparametric test will not work; you will not get anything significant. So, then you may still go and try a parametric test, but if you have like 10 or 15 samples and you are worried about the distribution, you might use a nonparametric test. So, again here you have to base it, you have based a choice on experience and the actual problem you are looking at.

(Refer Slide Time: 17:47).

Application of Hypothesis Testing: Finding Regulated Peptides

- Run experiment with groups of interest:
 - Case, Control
 - Baseline followed by multiple time points
 - Multiple disease states
 - Ex. Transgenic A, Transgenic B, Control
- Process data files
 - Calculate ratios relative to Control, Baseline or other “reference”
 - Log₂ transform if needed
 - Especially for ratios
 - Normalize
- Determine regulated items
 - Peptides, proteins, genes, etc.




So, like I have been mentioning that finding up or down regulated peptides is one of the applications of hypothesis testing. I think we have looked at all these, basically we the main thing here is, you need to log transform data, so it has a reasonable distribution and you can apply this to proteins, peptides, genes, anything you are interested in.

(Refer Slide Time: 18:13).

Finding Differential Peptides = Hypothesis Testing

- Case vs. Control: **one-sample t-test**
 - Is the log (case / control) ratio statistically different from 0?
- Comparing conditions: **two-sample t-test**
 - Is log (A/control) ratio statistically different from log (B/control)?
- Multiple Group Comparison:
 - Are **any** of the log (group_{*i*} / reference) ratios statistically different from 0?
 - *i* = 1, 2, ..., *k*
 - *k* = total number of groups

F-test or longitudinal analysis

 **BROAD INSTITUTE** Proteomics & Biomarker Discovery

So, I have some examples of different types of scenarios and what kind of tests you would use. So, you have a case versus control study where you want to know you have taken the log of the ratio of case to control, the logarithm of that ratio and you want to see which proteins are statistically different. So, in this case, you have a single ratio you have measured for all the patients and you have to come up with the measure of whether that protein was different in the cases versus controls.

The second one is comparing multiple conditions. So, you have condition A and condition B and you also have a control. So, this is the situation we would encounter in like the TMT study. So, the control would be the reference and the A and B would be like breast cancer subtypes or whatever. So, you have log ratios to some common control in our case, it was the reference. But if you had a normal sample and you want to see how things are different in basal versus luminal cancer with respect to normal, the controls could be the normal samples and A would be basal and B would be luminal.

So, you can set this up; however, you want based on the study you are looking at and the experiment you are doing. The other option is multiple group comparison, so I have 50 groups for breast cancer. Can I find a marker that is different in any one of those groups? So, that is a multi group problem and so for each of these groups you would do a different kind of test. So, for this you would do a one sample t-test. So, the t-test is one way of checking whether the mean of two things is different or not.



In this case, you are checking whether the mean is different from 0 the log ratio is different from 0. If it is different from 0, then it is either up in the case or up in the control depending on whether the ratio is positive or negative, the log ratio is positive or negative. If you are comparing two conditions, you would have a two sample t-test for multi group comparison. There is this thing called a F-test, but if the multiple groups arise from something like a time series. So, suppose you have the same sample, you measure at time 0 1 2 3 4 hours or days or whatever then the samples are related they are not independent. So, it is one thing to measure five different breast cancer subtypes, because they are completely independent samples, but if you took the same sample and measured it 5 times at different time points and you had 50 replicates of this. So, you have 50 different mice and for each mouse you measure something at time 0 1 2 3 and 4.

So, now you have 5 groups and many different samples, but the groups are all related, because they were measured on the same thing, so they are correlated. And so in that case you have to do this thing called a longitudinal analysis to take into account that the measurements or the groups are related. And so, that requires a slightly different way you analyze the data and F-test; a straightforward F-test is only if your groups are independent.

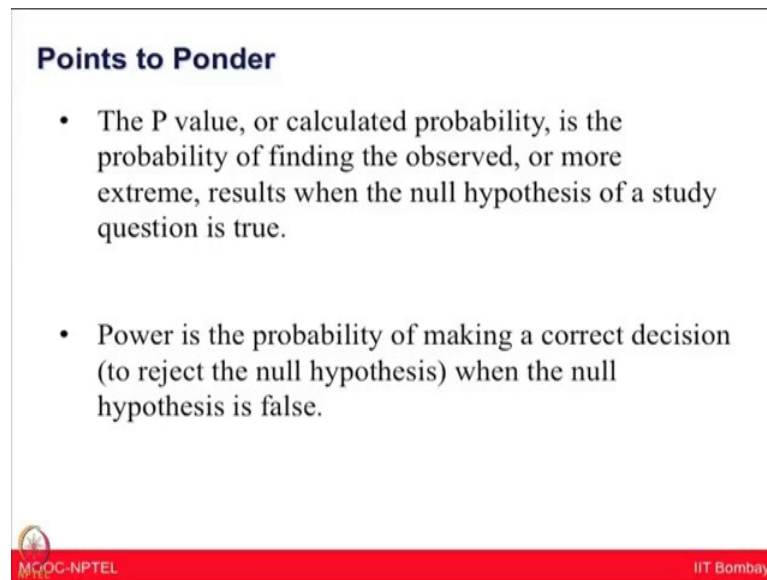
(Refer Slide Time: 21:25).

Points to Ponder

- Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data.
- In Hypothesis testing, if the significance value of the test is greater than the predetermined significance level, then we accept the null hypothesis.

 MOC-NPTEL  IIT Bombay

(Refer Slide Time: 21:37).



Points to Ponder

- The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis of a study question is true.
- Power is the probability of making a correct decision (to reject the null hypothesis) when the null hypothesis is false.

NPTEL IIT Bombay

I hope today's lecture was helpful for you to learn that making decisions in a statistical analysis include whether we should accept the null hypothesis or reject it. We understood that is hypothesis testing if the significance value of the test is greater than the predetermined significance level, then we accept the null hypothesis. But if the significance value is less than the predetermined value, then we should reject the null hypothesis.

Finally we understood how P-value and power of test is important for hypothesis testing. Power is the probability of making a correct decision to reject the null hypothesis or when the null hypothesis is false. The P-value or the calculated probability is the probability of finding the observed or more extreme results when the null hypothesis of a study in question is true.

Next session, we are going to have a hands on session about the software Protigy which will give you an idea how to analyze your complex mass spectrometry datasets and then use many of the essential consideration which we have talked in the last few lectures.

Thank you.