Research Methods in Health Promotion
Dr. Arista Lahiri
Dr. B.C. Roy Multi-Speciality Medical Research Centre,
Indian Institute of Technology Kharagpur
Week 11
Lecture 52: Quantitative analytical methods (Part II)

Hello and welcome back. We were discussing the quantitative techniques that we utilize when we are analyzing our health promotion research data. And in this lecture we will continue our discussion on that same topic. We left our discussion in the last lecture with the idea of bivariate and multivariate techniques. In this particular lecture we will be focusing mostly on the different statistical techniques that we utilize when we are actually implementing the analysis procedure. And we will also we will also try to correlate our idea behind bivariate and multivariate techniques and the statistical techniques that we discuss in this lecture.

So, the concepts that we will be covering how do we choose the statistical test appropriately. We shall discuss in brief about the t test and the chi-square test as the tests of association between variables. And also we will discuss about correlation and regression as the tests or the techniques that give us you know the measurement of association the or the strength of association as we call it. So, the examples of some commonly used statistical tests.

This is a very interesting table and I think you can get a table of similar kind if you even Google it because this is needed when actually you are going to implement your data analysis process. But we have said that whenever we are preparing for our health promotion research we have to be a priory sure about what data analysis technique we are going to implement. That is why whenever we are designing our study based on the variables and operationalization of of those constructs and concepts the measurements that we are utilizing we have to understand which type of analysis we can really perform. And we have to understand which types of test we have to perform that different types of test that you can decide this table will help you in that in making that decision. Now, we shall not be going into details of this test because these states these statistical tests they are not really part of this particular course rather we will give you an overview on which test to use you can perform all these tests utilizing the standard statistical softwares.

What you need to know is what interpretation you have to make from the results of these tests. So, those are the areas that we will be discussing we will mostly focus on the basic principles of test where you can utilize these tests and what interpretation you can get. Now, see in this in this slide what you can understand is we have you know differentiated ah the different tests in terms of the levels of measurement. See Noir Nominal Ordinal Interval or Ratio Scale for interval ratio scale we have similar kind of tests for ordinal scale we have a similar kind of test and for nominal we also have a similar kind of test. Now see during ah the

first lecture what we ah what we noted was that the measurements that we do and the test in the choice of test it depends on the distribution of our data.

If a data is normally distributed then we go for parametric test, if it is somewhat skewed or not normally distributed then we go for non parametric test. We also stated that the commonly used tests like t test, z test these are all parametric test where the assumption of normal distribution of data is maintained, but where the data is not normally distributed under non parametric assumptions there are different tests you will get some names of these tests like the Mann Whitney U test, Wilcoxon test and most commonly the chi square test obviously, it is also a non parametric test. Now, our our question is which I mean when to use which kind of a test. So, when you have two independent groups that means, you are comparing the behavior say among a group say among males and among females. So, the males and females are your two independent groups that means, you it now depends on the level of measurement of the outcome variable.

Say if you have measured say the behavior behavior in a continuous scale. Now, behavior how you have measured you have measured from say 0 to 100 scale this is a continuous thing and usually there are chances this if you if you get a multiple continuous variable the distribution it will be more or less normally distributed and and based on that for interval ratio scale what you can do if you want to compare this behavior scored among male and female you can go for independent samples t test or what is commonly called as unpaired t test. Now, if it is not like 0 to 100 if it is simply certain ordinal measurements like what we get in an like at scale 1 2 3 4 5 like this or simply 1 2 3 like this then it is not normally distributed and neither the data that we get is is interval ratio data. In that situation the data that we get it is ordinal data and for that you have to use the Mann Whitney U test. These are all the standard nonparametric test and in any statistical software that you can use you will get all the these tests built in those softwares and you can simply run those tests by selecting the outcome variable that you want to test comparing with the comparison group ok.

And if the behavior it is measured simply like yes no it is not ordered in fact, then you have to perform the chi-square test it is nominal. So, for nominal to independent groups you do the chi-square test. Now if you have to dependent groups that means, they are somewhat related amongst themselves. For example, if you have a matched pair case control study now the matching ensures that the two groups are related to each other ok. So, in that situation what happens is for these nominal kind of situation this yes no situation you use the McNemar test.

It utilizes in this kind of a table we will come to this table as the contingency table discussion ah in in next slides. Now if you have this kind of a table these are your cells A B C D these are say the this is the outcome and this is the exposure and you have matched pairs what the McNemar test does McNemar test considers these discordant I mean these discordant pairs ok these are called the discordant pairs that is how McNemar test ah helps in identifying if these two variables are related to each other or not. Another thing is that the McNemar test is

basically in principle is a modification and a stringent modification of the chi-square test. The stringency or the as the statisticians call it the conservativeness in the test comes because it accounts for the relationship between these two groups. Because in most of the statistical techniques or in most of the analytical principles we assume that the different groups that we are testing they should not be related to each other because if they are related to each other because of that correlation your interpretation or simply the assessment of these distributions can vary that is why McNemar test it it creates a more conservative effort.

For ordinal level variable like this for two dependent groups you have the Wilcoxon test Wilcoxon rank test and for the interval or ratio level data like this you have the pair t test. Now for more than two independent groups you have three four groups you still can perform the chi-square test if your ah if your outcome is nominal like this, but in that scenario it will be called a chi-square for trend you are observing the trend of ah of association between the outcome variable and also the predictor variables ok. If it is ordinal then you can use the Kruskal-Wallis test and if it is interval or ratio scale you can utilize the ANOVA test. Now for ordinal level of measurement if your ordinal level is extending to say more than six or seven ah seven I mean points you can utilize the ANOVA test as well because in that situation what happens is that your measurement it is considered as an interval or ratio measurement ok it is it becomes a continuous measurement like this. If you have more than two dependent groups that means, again if you have different say three or four matched pairs whom you are comparing say the 1 is to 2 case control study that we devised we have matching of ah one case with two different controls.

So, here you may have three matched pairs 1, 2, 3, 3 different groups, but they are matched for nominal level Cochrane skew test for ordinal level like this the Friedman's ANOVA by rank. So, in all the ordinal tests basically the as the basic nature of the test is ranking them and for interval or ratio scale you can go for repeated measures ANOVA. This is again a multivariate technique ah and although you are basically testing between two variables, but the principles that come from the repeated measures ANOVA technique ok. So, now we move on to the discussion of the basic tests that we mostly utilize in our health promotion research. Remember not all the tests that we discussed we will we will be discussing we do not have that kind of a scope for this course we just have given you the overview of how to choose your test you can ah memorize this table or keep this table handy whenever you are planning your analysis of the data.

So, what happens with the t test? A t test is basically testing the continuous variables that you have as the outcome in terms of any association with the predictor that you have considered or the independent variable that you have considered in your in your study. It is basically the common goal of the observation research in health promotion is to compare the subgroups of a sample that is defined by a grouping variable for example, gender may be a grouping variables where the subgroups will be male and female with respect to second variable second variable may be behavioral intention that is your outcome variable ok. So, what t test does is t test compares the means of the of the means course of the behavioral intention among males

with that of the females. Now the t test can be a paired t test and an unpaired t test. Now the pairing and unpairing it depends on whether we have the independent groups or we have the dependent groups ok.

So, when we have a pre post comparison like BP measurement before an intervention say your health promotion intervention is here and you are doing a BP measurement before and again a before BP measurement after ok. Sorry and what happens here is this is you have measured the mean you have again measured the mean and you are comparing these two means. See essentially these are the measurements from the same individuals that is why they are linked that is why we have to use the paired t test over here. But if you are comparing the you know the BP measurements from the male and the female or the behavioral intention among the males and the females of your target population then they are not typically related or at least not how you have designed them to be related. You can easily perform the usual unpaired t test to find out any difference in the mean scores between the two groups.

So, t test always gives you an inference about the mean scores, ok. The typical outlook or the output from the different softwares are like this. Here you have the different measurements what you have got here you have two lines like whether the equal variances were assumed whether equal variances were not assumed here you have one p value for the test of equal variance if you get the p value to be significant then you see the you consider that there are the null is true that means, I mean there are significant differences. So based on the your interpretation of whether the there is variance difference in variance between the two groups or not you have to infer the output of the t test which is typically presented after this kind of a gap. Here you have the mean difference all the confidence intervals for the mean difference and the p value always remember to infer the p value from the correct row based on whether the variances were different or not and this p value you have to only report ok.

Next is the chi square test typically this is the this table is called a contingency table see here we discuss that this kind of a table will be coming to later on and this kind of a table it is essential for data analysis in health promotion in epidemiology and as a whole for biostatistical data analysis. See here in contingency table the columns here you have the outcome lung cancer no lung cancer and in the rows here you have the predictor variables smoking or non smoking without any doubt the outcome variables over here they are basically nominal variables and that is why here we should be doing the chi square test. How do we do the chi square test? Through the chi square test we basically assume I means we basically assess whether the predictor variable is related to the outcome variable or not remember this is only the test of association ok. Now this total over here this is the total we call it a grand total. So, the total of the columns and the total of the rows they should be equal that concludes a contingency table it creates a formal contingency table.

Now here you also have to understand one more concept that is called the degrees of freedom degrees of freedom to put it simply in how many ways you can put a particular data if you

have fixed the rows and the column values because see in a contingency table each column and row has a total and that has to be fixed. So, based on that you see over here you can put a single data in in a single way see if you have fixed this number and you know that non smoking people they are 75 you can only put 35 over here you do not have chance to vary this. So, the the common formula to identify the degree of freedom is the number of rows minus 1 multiplied by column number of columns minus 1. So, here number of rows are 2 number of columns are again 2 we do not include the total in this calculation. So, 2 minus 1 1 multiplied by 2 minus 1 1 that gives you 1 that is why here the degrees of freedom is 1 that means, you can put this value only in a single way or any other value as a matter of fact.

So, the when you are representing the you know sorry whenever you you are representing your chi square table data be very sure to present the degree of freedom and the computed chi square value because that that is usually the general requirement whenever you are going to publish and that gives an idea to the reader that ok the p value this is how the p value was designed and and the p value whether it is significant or not. Next is the discussion of measurement of strength of association. See the first ah discussion topic would be a correlation. How do we ah what do we mean by correlation? Basically it means whether two variables co vary. See in previous lectures we discussed that there can be different types of relationships.

Now here the relationships are like this ok. So, the basic idea behind having this kind of a relationship is to show you that with the increase in x the value of y is also increasing. So, that means, x and y are varying simultaneously depending on x y is changing this is called co vary. Now how do we measure how these variables are co varying and if we can measure the extent of this co variation like for example, delta x marks the change in x and say delta y marks the change in y. If we can measure the the ratio of say delta y by delta x we can say that with one unit of change in x y is changing by this much.

This gives you a measurement of how much change is occurring or how much y is co varying with x. This particular term how much the effect is occurring is called the effect size and that is measured in this case with the correlation coefficient. It is basically a statistical index of the degree to which two variables are associated or related. So, you measure the degree over here. Now usually we implement the strategy of Pearson's correlation when we have the ratio or interval level data like where we have you know what we discussed like the scores 0 to 100 scores it can take behavioral intention or as we have been discussing lately with the self efficacy beliefs and also the PVC we can in in an arbitrary way try to compare the relationship between PVC and self efficacy.

Say PVC is measured again in this continuum and self efficacy beliefs are also measured in this continuum. Then these two scores they become continuous and we can consider them as the as norm maybe as a normally distributed and as part of the interval or sorry the ratio scale. Then the correlation technique or measurement of correlation coefficient will give you a fair bit of idea of the effect size. That means, how much self efficacy beliefs can vary if the

perceived behavioral control varies, but this is only an arbitrary example. You have to choose these variables based on your hypothesis and also the research question keeping in mind the biological possibility as well.

Usually the correlation coefficient it is presented by R you know the small r it represents the correlation coefficient and the larger r, r squared it represents the index of determination. The small r the value ranges from minus 1 to plus 1 and the minus 1 or the plus 1 they represent perfect correlation. So, in this scenario plus 1 means a I mean it is the it is the highest it is the perfect correlation between these two variables and minus 1 means it is the inverse relationship and again minus 1 means it is the perfectly possible relationship. That means, with increase in x y is decreasing that is the highest possible correlation that two variables can have. So, as you can understand from this discussion that the correlation analysis is basically bivariate technique because you typically have two variables one of them will be your outcome another will be a predictor variable.

The Pearson correlation coefficient typically represented by R is used when we have this kind of a continuous data. But what do we do if we have certain ordinal or nominal measurements like say in this case the ordinal or nominal tables we did not mention about the correlation or how to measure the strength of association. So, for that you also have some different kinds of correlation measurement techniques like this Spearman's rank correlation technique you have Spearman's rho in that scenario it is not r anymore it is rho, but still the value the range of the values it remains the same minus 1 to plus 1. So, the correlation always whenever you are trying to implement the correlation analysis remember to check what type of variables are you dealing with you may deal with two continuous variables you may deal with one continuous and one see this kind of a nominal ordinal variable like this. Be very careful to choose your technique of measurement of correlation coefficient, but mostly for continuous variables or continuous this core kind of variables we use the Pearson product product moment correlation that is the most commonly used correlation technique.

Next we come to the concept of linear regression from this slide onwards we will be discussing about the regression techniques that we have in our health promotion research. Regression techniques can be manifold, but linear and logistic regressions are mostly the the most commonly used regression techniques. The advantage of regression is through correlation we could not study the effect of many other variables and how they are co-varying with the outcome or how the outcome is covering with the other independent variables, but through these regression techniques we can adjust for the different other variables that can have an effect on the outcome on the final outcome variable that that is why the regression techniques are often multivariate techniques. Then the multiple variables instead of only the outcome and a single predictor when multiple variables involved we typically call it a multiple regression. Now what is a linear regression? Here your variable if you consider it in a two dimensional way and we are regressing regressing y in terms of x again it is as simple as the correlation we have the same line this is called the regression line and you have the same data our target is to

fit the line in so, in such a way that the differences from the these points they are the some of the average of these differences they are minimized ok.

That indicates the best fit the best fit line. So, the regression basically identifies the best fit line, but linear regression means that the x and the y they are on a linear in a continuous scale like behavior and intention in terms of 1 to 100 for knowledge in terms of say 1 to 50 like this. So, linear regression it shows a linear relationship and it requires the outcome variable to be continuous in nature. So, in multiple linear regression you can use it to judge the collective strength of all the x variables in explaining the variance in y. So, whenever we are trying to minimize this distance it means that you are minimizing you are trying to explain the various variability or the changes the this the gaps from the regression line to the exact point they represent the variation from the regression line.

So, you are trying to basically explain the variation from the regression line with the help of a equation and with this best fit line. Whenever you are doing a multiple linear regression you will be having a different dimension and see this is another dimension and it will be a whole n dimensional or multiple multidimensional data and in that situation the regression line will somewhere be like this. You have to conceptualize this in a 3 dimensional or 4 dimensional way and that line in that scenario it compiles all the x variables and it understands it it shows what is the effect of all these x variables on y. How do we get that? We simply get it through the beta coefficient that we get from our output or some software it demonstrates it is as simple b. And remember it is always better to use the standardized coefficient instead of the unstandardized coefficient, but some journals may ask you to report the unstandardized coefficients as well.

These are just the terms that you have to watch out for when you are performing your analysis for the health promotion data and as mentioned in the previous slide also any statistical software can perform these kind of analysis these are simple analysis. Another term called multicollinearity we have to be very careful about often the multicollinearity measurements are also included in the softwares and you have to assess for the multicollinearity because what is this? This is the effect manifested on correlation among the different x variables. So, these are the in these different dimensions these x variables or the predictor variables they are related among themselves. Say for example, expenditure and income they may be related and you may have collinearity between expenditure and income. So, if in a regression model you try to explain the change in certain behavior or performance of a certain behavior in terms of both income and expenditure and you suddenly see that income and expenditure related what happens is the relationship the true relationship between these variables they will be distorted and it may so happen that you will fail to document any statistically significant effect of these variables on the particular behavior.

So, that is why it is very essential to understand the presence of any multicollinearity or not. If multicollinearity is present then you have to exclude one of the variables or two of the

variables like this that will depend on you on what inferences you really want to have and what are the biological plausibility parameters of this regression equation. Basically the linear regression it serves several purposes, it tests the nature of linear relationship between the x variables and the outcome y, it gives you an idea which x variables or which predictor variables actually have a statistically significant relationship with y and which do not and it is also gives the strength of that relationship. That means, with change in x how much change in y is there that means, if we say if we consider it in terms of a health promotion research what we can get is if we implement that particular you know that particular intervention how the behavior change will occur. It is as simple as that there that particular strength is depicted by linear regression and remember the linear regression can be done when the data the that we are do it that way analyzing a continuous in nature.

It can also form a prediction equation see from this line we have data from this portion and we have developed this line, but if we have in future data on x in this segment this line will give you or an estimation of where what will be the y value in future or in other positions. So, this is called a prediction. So, linear regression gives you a prediction equation as well from where you can predict the values or the predict the change in the outcome variable and that is one of the very much essential analytical techniques we utilize in our health promotion research. Next is the logistic regression this is again a very basic technique of taking inferences also usually if you include more than two variables for logistic regression analysis it becomes a multivariate logistic regression and this logistic regression is different from linear regression in such a way that the linear regression is mostly a parametric test and the logistic regression is mostly a nonparametric test why because in logistic regression the outcome variable typically you have the outcome variable it is dichotomous simply like yes or no did they perform this behavior yes or no and you predict the performance or you understand the performance of this behavior through different other predictor variables that may be dichotomous that may be polychotomous like this or in fact, ordered or in fact, continuous. The core thing is the outcome should be yes or no typically when you are performing these kind of analysis in the statistical softwares you have to mention code these variables like 0 and 1 thus software basically you know models 1 against 0 and when you get the output from the software what you have to observe is again the exponentiated beta value.

The exponentiated beta value gives you the odds ratio as we have mentioned over here and this odds ratio is the interpretation that you get from any logistic regression analysis. That means, the odds of occurring the event occurring the incidence against odds of not occurring the incidence for say particular exposure factors. For example, in a smoking example what are the odds of smokers getting lung cancer that can be answered through logistic regression analysis. Ordinal regression and nominal regression this happens when the outcome variable is not dichotomous ordinal means the outcome variable is ordered in nature and nominal regression means you have multiple categories in the nominal variable. For example, in the nominal variable you may have interested, engaged and then both.

Now these are the categories see these categories of the nominal variable you cannot club them into two. In this situation what you do you go for the nominal regression again the interpretation will remain the same you have to look for the exponentiated beta. Here always remember to mention the goodness of its statistic in all regression analysis you have to mention whether the model was effective or not through the goodness of its statistic. For linear regression while you have the different techniques for example, the variance explained the r squared like this in logistic regression that r square typically does not is not as similar in interpretation as we get in the linear regression. Here we have some more measures like the log likelihood test minus 2 log likelihood ratio you have to mention what is the minus 2 log likelihood ratio.

If there is an improvement from any previous model you have developed or not all of these you will get from the output of a statistical software, but be very sure for logistic regression to mention the Hosmer Lemmichaux statistic whether the test of sphericity whether this the model was spherical or not it is always recommended that for a slow district regression model it should not be spherical ok. So, all of these you can get from the output of the analytical software's. Finally we come to the concept of count regression techniques here the data is continuous, but it is countable that means, 1, 2, 3 like this number of individuals like this. The advantage of using a count regression over logistic regression is that it typically gives you the relative risk or the or the relative risk ratio the software some software it provides you as the incidence rate ratio like this. So, that is why we tend to provide you know we tend to mention count regression in certain longitudinal studies because through count regression we can get the relative risk what is the of that in of that longitudinal study.

But if we have implemented the logistic regression for analysis of the data from that longitudinal study we could have only get the odds ratio and we all know that there is a relationship between odds ratio and the relative risk although the relative risk and odds ratio may not be same. So, always it is better if you have a longitudinal data you should go for the count regression typically the Poisson model when you use it the incidence rate ratio that is provided in the output of your table, it is the relative risk that you get. So, even if you have a yes no response you still can perform the count regression to get the relative risk, but if you have a count for example, the incidence of occurrence of certain accidents 1, 2, 3, 4, 5 and against that you are performing this Poisson model to understand what are the factors that led to the accident. The interpretation that will get that you will get is the relative risk of the different factors for example, not wearing helmet you will interpret it like this a relative risk of not wearing helmet is this much for say for having one more episode because in count what you have 1, 2, 3, 4, 5 means it models again having one more episode of the of the events ok. So, count regression is a bit advanced yet a very useful technique when you are analyzing your health promotion data.

Again this is a table that you can understand how to test the association and the the the strength of the association you can consider this table as well also the first table that we mentioned. So, in conclusion we discussed the different choice of our tests how we have to

choose we have to look for the dependent variable what is the nature of the dependent variable and what is the nature of the independent variable, what are the number of groups that we have 2 groups more than 2 groups like this based on that we have to choose whether t test we have to observe whether the groups are dependent on each other whether they are related or not based on that we can choose on say for example, the paired t test or unpaired t test like this. We discussed correlation and regression from the point of view of assessment of the strength of relationship correlation gives you typically the Pearson correlation gives you the small r which gives you how much y changes with the with x the that means, the outcome changes with the predictor, but the regression analysis it helps you in in understanding the effect of all the different predictor variables on the outcome that is why simple multivariate regression techniques is useful in health promotion research because you have an adjusted estimate of what exactly effect is there on y. Similarly, we may discuss that logistic regression is used for dichotomous variables when the dichotomous variables are for the outcome variables there are certain deviations of logistic from logistic regression when the outcome variable is ordinal in nature we use ordinal regression if we have nominal variable with different categories more than 2 categories typically we use the nominal regression technique and lastly we discussed that the count regression technique is very much effective in providing an information on the relative risk typically where we have the longitudinal values it is better to perform the count regression technique. So, that we get the relative risk instead of only the odds ratio that typically the logistic regression gives.

So, these are the resources for this particular lecture and in the next lecture what we will discuss is general outline on how to go about the analysis of our health promotion research both for observational and the interventional data now that you have a cross overview of what are the different analytical techniques and you know that these techniques can be implemented through any analytical software. So, that is all for this lecture. Thank you.