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 53: Quantitative analytical methods (Part III)

Music Hi there. So, we were discussing regarding the quantitative methods for analyzing our health promotion research data. Now, this is the part 3 of our quantitative analytical methods discussion. In this discussion, we will focus on the basic principles of data analysis in our observational experimental research, then the framework of data analysis for both observational and experimental research. We discussed what we understand by effect size this is a very important concept and also briefly I will give you an idea regarding the different statistical software that we can use for quantitative data analysis. Now regarding the principles of quantitative analysis.

Basically these are rather good practices then to be considered as principles. One of the good practices plan the analysis beforehand that means, at the protocol phase we have mentioned that whenever you are proposing the research, whenever you are developing your protocol it is obviously, always better to mention all the issues that you are going to take care of and also mention how you are going to collect data, how you are going to analyze the different variables and also what are the different techniques that we are going to do. So, plan the analysis beforehand. Next you have to identify the variables and their measurements.

Basically what happens is once you have this, once you have identified the different variables that means, how exactly you are operationalizing the concept into a construct that that helps you to identify the different variables. Then the diff through the different variables you are basically measuring a particular constructs and once you have done that now you know how you are going to analyze the different variables to measure the particular construct and also how the particular construct that you measured will ultimately influence the outcome that you also have planned. Analysis should correspond with the research question that means, when when when you have a research question to find say for example, the predictors what are the different predictors of a certain good behavior for example, of healthy dietary practices. Now here your research question deals with the predictors of good dietary practices. Here your analysis should directly focus on finding the predictors you may consider using the chi square test test for association like this.

Now if you do something that is not really focusing on finding the predictions for example, you are you know you are simply doing a descriptive analysis or you are simply describing the data in terms of mean median or central tendencies and the measures of dispersion. Now what happens here the typical analysis it is not in line with the research question it is not able to give you the results on whether the particular factors like ah you know ah particular knowledge

is influencing in the good dietary behavior or not it will not give you, but if you have if you perform the analysis through certain association tables it will give you certain ah result. So, your analysis should always correspond with the research question again this should be considered beforehand before actually performing the the analysis and at the inception phase. The use of unplanned tests is thus discouraged. So, if you have planned the whole thing ah when you have actually operationalize the constructs and developed your research proposal it is obvious that you will not be needing any sudden any suddenly any other tests to find out ah the different ah results.

That is why also there is there is one reason that if you have certain unplanned testing ah say you have planned for a chi square test, but now you decide that no I will I would not want to test these variables in in dichotomous form I want them in continuous form and I will be I will be testing through a t test. What happens is the samples as that you have calculated for the chi square test is not may not be same as that for the t test. So, you need an additional power analysis and ah and usually ah it is not recommended to go for certain unplanned tests because it may not be having certain a statistical validity if you have not ah proposed it a priori. So, what about the statistical tests in experimental and observational research? In the previous lectures we discussed about the different ah different testings different statistical tests that we use for describing the data and also for different inferential parts. So, exactly what do we do? Here what we have to understand is there are certain subtle differences in analysis of experimental and observational research.

See if you remember the experimental research they are more of a definitive research there they will give you certain conclusive evidence, but the observational research they are more of you know finding the flaws and finding further research question kind of research. That means, the observational research they are more open and they give you more freedom you know to explore certain things. For that reason whenever you are analyzing the quantitative data from your observational research you may consider using certain unplanned tests. That means, once you have the data you ideally you previously thought that ok I will perform t test because ah this is a continuous measurement say this is a measurement behavior B 1 is the measurement and you have two groups say male and female. Now you you planned that ok I will perform a t test between these two groups to find out the differences in mean and standard deviation.

Now here since this is only an observational research where you are generating your hypothesis or generating your further research question for experimental research, but you can do is you find that no the t test results they are not very much similar. So, what you can do you can consider dichotomize the B 1 based on a different different cutoffs say this is a scale of 1 to 100. Now you can consider a cutoff of ah more than equals to 50 and less than 50. Now based on this dichotomous data you can consider one contingency table ok and based on this contingency table you can find out association. It may so happen that this unplanned chi square analysis or this unplanned association analysis this will give you certain important inputs it may be the p value may be may become significant.

Obviously you will do the the the power analysis for this unplanned test to prove that this test it has a sufficient power to have a strong conclusions to draw some strong conclusions, but this p value will give you certain insights on how to frame your next research question or how to move forward with the experimental research. It may so happen that it will give you an idea that there are differences in gender, but that difference is not homogenously distributed there are certain other distributions across different ranges of the value of this behavior. There may be certain moderation of this gender ah as gender variable on this behavior. Now this is what you get from the different statistical tests from the observational data and you keep this in mind to propose your experimental research because remember the experimental research is more of a confirmatory research. Here you definitely know these are the variables you definitely want to understand how these variables are affecting the outcome.

So, again this is a confirmatory thing. So, for that you really cannot deviate from your plan of analysis you cannot suddenly in a introduce another type of analysis because the samples is calculation the the data collection all of these have been designed based on how you are going to analyze and vice versa that means, based on all this your analytical method has been devised in the protocol phase. So, this is ah this is the the information that I wanted to give you that although it is always discouraged to into to to use unplanned tests, but for observational data since you are now generating the hypothesis generating research questions for your further experimental research you can introduce certain unplanned tests and you can you know keep on doing certain relevant tests obviously, to give you certain some more information from the data, but always remember you have to perform sufficient perform power analysis to show that the test that you have performed the unplanned test it is valid on this data ok. So, why basically do we need to you know undertake the the statistical tests in the observational research because the observational research is usually used to answer two distinct varieties of research questions. Now, most common example is is can be the research can be designed to simply predict why that is the correlates or predictors these are the typical things, but we mentioned that gender as a predictor of this behavior this is one way of looking at this question this is the this is the most commonly used observational design and often we use the cross sectional designs as well.

So, the other other variety is rather not that much common here the research begin with only a single x that means, say now you have this behavior and you want to understand whether this behavior is influencing certain other practices you have practice 1, practice 2, practice 3. See in this question this behavior is your predictor or explanatory variable and these different practices they are your outcome if you consider it in that way. You started with the single x variable the predictor variable or the x variable and then determine the relationship of this variable with multiple outcomes how it is related with P 1, P 2 or P 3 the different practices. Obviously, this kind of approach it helps you in addressing a very broad range of questions and it also helps you in understanding or raising a broad range of questions or a or different research questions a broad array of research question. But still this approach is not very commonly used because you know typically when we discuss regarding the mixed methods research I hope you have understood that for this kind of approach you have to first focus on

a particular behavior that you want to study and whether that particular behavior can really have interactions or can really influence certain other practices all these things you have to we have to prepare before actually going into analysis and also you have to think of what different variables outcomes we can think of based on that based on that behavior usually this kind of information you can get from the previous qualitative research.

So, again based on those data you can plan your your final analytical steps. See what happens over here is if you find that you have only P 1 and P 2 and no P 3 you do not get the behavior P 3 sufficiently then although you have planned for certain analysis with B 1 and P 3 you will not be able to do it rather in that scenario you have to find out how P 2 and P 1 may be correlated like this. So, what I am trying to tell you is you may you may encounter different scenario in the in this sort of quantitative analysis and for observational research you can really take up these different unplanned testings, but never for experimental research. I hope at this point this concept is very much clear because I have I am telling you repeatedly because this is what you have to be very much cautious when you are actually analyzing your own health promotion research data. Next we come to the discussion of the framework of our data analysis.

So, it is simple we discuss the description of data and then drawing inferences from it. So, it is basically like that first we describe the data how do we describe the data the key thing basically when we are you know using two group designs or certain experimental designs you have to assess the comparability between the two study groups. Remember what I have mentioned over here randomization does not ensure that the study groups are equivalent it only ensures that there is no systematic bias in the assignment of participants ah to the two conditions perhaps. Now when describing the data apart from plain and simple description of the different ah baseline characteristics for example, age gender that we typically do this assessment of comparability between the two study groups is also important. Although this part is typically relevant for ah the two group designs or multiple group designs or the randomized control trials, but for simple observational studies you also should consider if this is applicable or not say for example, a simple case control study in that scenario you have two groups although you have you have implemented matching, but you should go back and look whether matching was appropriately done or not whether there is any difference between the two groups regarding the different factors or not like this.

So, this is basically descriptions here you are generally describing your study population and you are not drawing any certain conclusions or inferences or ah certain other ah tests ok. Next is when you want to now make certain conclusions and inferences you have to understand the different types of different dependent variables. You have to understand whether you dependent variable that you have or the outcome variable that you have is categorical dependent variable, ordered categorical dependent variable, a continuous dependent variable or not. Categorical and continuous variable as you have understood from the last lecture we utilize usually the logistic regression models for the categorical variables categorical outcomes and we often follow the linear regression models for the continuous outcome. So, it is important here to understand what type of outcome variable we are planning in this scenario.

See usually in the RCTs or in the experimental research we know how we are going to ah analyze our outcome variable we know whether the outcome needs to be dichotomized or not as simple as that. So, based on that we already proposed that ok if the outcome is dichotomized that means, categorical binary categorical variable we can utilize simple binary logistic regression. If we find it that we need certain ordered categories in our in our outcome then ah we can utilize the ordinal regression technique again the ordinal logistic regression that we can use if they are simply the nominal ones although usually in health promotion research the outcome variables in health promotion research they are hardly the nominal variables rather we find the mostly the dichotomous and the ordered variable. So, if even if it it it was nominal one we could have utilized the nominal logistic regression and for continuous variable already mentioned we can use the linear linear regression, but the important thing over here is you have to specifically mention it. Next is the the very important reason why we often transform our continuous data into categorical data we may find a behavior in this continuum say 1 to 100, but instead of using it as a continuum we tend to transform it ah to say less than 50 and more than 50 more than equals to 50 into 2 groups why.

The first thing is if we simply use the linear regression it will give you the beta component or the b the the regression coefficient and if you perform a logistic regression it will give you an odds ratio see it is obviously, always better to interpret an odds ratio than interpret the regression coefficient this is one reason for understanding. Next the statistical reason of transforming this variable is sometimes this outcome variable it may violate the assumptions like the assumption of normality for performing the parametric test or ah say ah the these kind of assumptions. So, when the assumptions are violated what happens is you really the test it is not valid in that scenario you can convert them to certain dichotomous or certain ordered categories and you can perform the ah non parametric tests. So, this is again an important understanding of the outcome variables what is the outcome variable whether you need any transformation or not and what is the validity of the transformation if you are performing it. Next is you have to select the appropriate statistical technique what we have already discussed with the discussion of the appropriate ah you know dependent variables where we utilize the logistic regression binary logistic regression ordered ah ordinal logistic regression like this.

So, here ah basically what we have to understand is that that data analysis it is a process it is a continuous process. Now at the each juncture in the data analysis process the investigator will be faced with making the decisions regarding statistical technique that is most appropriate for analyzing the type of dependent variable collected. So, again the the emphasis is the type of dependent variable that we have already understood in perhaps in step 2. Now this decision mapping on which kind of analysis I need to perform it is really based on the on the types of data that you need to analyze and the qualities and characteristics of those data. That means, the matrix of measurement that we discussed long back in week 4 again that comes into play how we have measured the data and what exactly do we want to understand from that data.

Say this the thing with the B and O R this is a classical example of choosing a particular statistical test. If we want to understand the odds ratio of say a better behavior a higher score then it is always our customary to go for the logistic regression and not for the linear regression because this will make more sense. These are the different attributes for the data that you have to consider when you are deciding on the statistical test that you want to conduct that is what we have mentioned it usually directly depend on the type of data represented by the dependent variables. So, now briefly we will go through the different techniques that will be using and what do we get from implementing these different techniques. So, multivariable model with categorical dependent variable or you can have you know not multivariable bivariable model as well, but when you have categorical dependent variable logistic regressions like binary logistic, ordinal logistic and multinomial or nominal logistic regression you can perform based on what is a dependent variable.

Usually in health promotion research we get these two types of ah outcome variables commonly. What your logistic regression will give you it will give you an odds ratio of the health promotion effect or the health in effect of the health promotion intervention ok. Say if you have proposed this intervention I you have given this plus minus now you know about the outcome plus minus. So, here you have a contingency table where this part this part is outcome positive after giving intervention here you do not have any outcome without the intervention and these are the discordant pairs like this you will have certain contingency tables. Now from here you can calculate the odds ratio it is simply we typically call them as a d by b c that means, the concordant pairs multiplied divided by the discordant pairs multiplied.

So, this will give you the odds of having say the positive ah behavior or positive outcome ah among the among those who got the intervention compared to those who did not get the intervention. So, this is this makes more sense you can really understand how much this intervention is affecting in terms of performance of the good behavior like this. Next is you have multiple follow up assessments typically when you have longitudinal ah follow up studies or the the different experimental research where you are measuring the individuals or the participants at more than one time point. So, there you have to keep in mind you will encounter two kinds of variable one is time invariant that does not change with time another one is time variant that change over time. Now a person's gender usually typically considering it does not change with time ah it is considered time invariant variable.

However, see the socio economic status of an individual or the monthly income of an individual it may change over time you can consider it as an time variant variable. Again it depends on how you have selected the variables at the protocol phase and how you are going to ah handle the variables ah when you are collecting the data it all depends on on on these things whether you have to classify that variable as time variant or invariant ok. Here the important thing is the generalized estimating equations this is another new term that I am going to introduce to you in this in this lecture this is the recommended strategy. Why because of the repeated measurement what happens is the measurements that you have they are usually not uncorrelated rather for example, if I I have I am being tested twice my basic attributes will

remain the same when I am being tested for the second time. Now because of that the responses that I get from the two testings they will be correlated.

Now this correlation we often call for this we have a statistical clustering the statistical clustering effect you really cannot overcome through simple logistic or linear regressions and for that for this effect of time remember this is the effect of time you need to implement the ah the generalized estimating equations model this considers the clustering effect due to the time or due to repeated measurements. Next analysis with a continuous dependent variable you have to choose whether you can perform parametric versus non parametric tests typically in the bivariate phase also ah in the in the multivariate phase what you can do is you have to consider whether you can go for a simple Pearson's correlation or not or whether you have you can actually implement the the linear regression techniques or not all these you have to consider ok. But remember from this you get the beta if you implement the linear regression only you get the beta standardized coefficient regression coefficient ah it it is interpreted in terms of how much one unit change in the predictor variable will bring change in the outcome like this ok. So, it is although it is also easy, but somewhat you know come more ah it is easier for to for everybody to comprehend the odds ratio as compared to this ah standardized ah regression coefficient. Next you have the multivariable models with continuous dependent variables as we were mentioning the regressions also you can have the analysis of covariance the ANCOVA or the multiple regression technique multiple linear regression technique is perhaps more common it gives you the beta.

And remember for ah the parametric test like t test it will give you the mean difference ok the mean difference as ah the outcome and you can comprehend based on ah the bivariate test if you get the mean difference how far those two groups are say the mean of a particular behavior among the males is this this is male and this is female. Now in this group and in this group if you have the difference it is the mean difference that you have and you can easily understand how distinct these two groups are or you can simply make an inference from there and for regression we also mentioned about the beta. Now this brings us to the penultimate topic of our discussion regarding the effect size in the quantitative analysis. So, what do we mean by effect size? Effect size means whether you know the strength of association basically whether the the measurement that you are doing it focuses or it shows how well the x variable the predictor and the y variable how well they are related. One you have measured the association we discussed and we also mentioned that through the different regression techniques and often sometimes through also the correlation we understand what is the how well these two things are related.

So, that is basically given by the term effect size this is one of the very common terms used in epidemiological research and also in in health promotion research. So, this gives you the magnitude of the difference in the two outcomes based on the the intervention may be that you have given. Now there are a few measures of the effect size we consider typically the categorical outcomes in this in this particular slide. Remember you can have the percentage or proportional difference because you have one outcome say in the intervention group and in the

intervention negative those you have given intervention not given intervention you have certain outcome positive out of total and in this group also you have certain outcome positive out of total what you can do you can have a percentage in these two groups and you can just have a simply you can have a difference between these percentages and that will give you a percentage difference. The more common used more commonly used method is relative risk and it is better understandable for the general audience also basically it is the ratio of the probability or as we commonly call it as the risk of an events occurring that means, developing a disease or sustaining an accident like this in an exposed group to the probability of the event occurring in an comparison unexposed comparison or unexposed group like this.

So, it is somewhat similar to the understanding of the odds ratio, but here you are actually measuring the risk in odds ratio taking the ratio of the odds in the two groups here you are taking the ratio of the risk or the probability of occurrence of the disease in the two groups ok and always remember that it is always better to measure the relative risk if you if you can because it gives you a better measurement whenever the the events they are more frequent usually the odds ratio they tend to exaggerate the actual effect that means, they move away from the null, but relative risk it it usually minimizes the exaggeration that the odds ratio does and there was therefore, relative risk is a better measure of effect size when you are doing your study with a very common outcome. Also one thing you have to remember that you have to report the confidence intervals when you are reporting all these outcomes and it is essential. For the continuous outcome what you can have you can have relative difference as we mentioned about the mean difference from mean 1 to mean 2 say this is the mean difference also what you can do is you can have different you know different populations these are the different values and again these are the different values in this group 2 see this is group 1 this is group 2 you can have a difference between these two groups and you can you can model based on the difference. Very common use in this scenario is the use of percent relative difference say for example, this is not G 1 and G 2 only say this is time 1 and this is time 2 you have measured the same individuals. Now, what you want to understand is if you take the difference the relative difference between this individual 1.

So, the behavior say behavior measurement over here is 50 and over here the measurement of the behavior is 75 the higher the better that means, after this certain time point for this individual 1 the behavior has improved. So, what we can do through what we can measure over here we can measure the difference difference is 25 and the what is the percentage of relative difference 25 say out of 50. So, you have a 50 percent relative difference for this individual. Now, I am I have given this example in terms of one individual, but exactly when we are analyzing the data it becomes you know a cumulative effect ok. So, but the basic idea of calculation of this percent relative difference and interpretation of this percent relative difference although remains the same.

Next we come to the issue of subgroup analysis. This is again a very important issue here in subgroup analysis basically what we do ah you take more and more groups and you analyze the different ah the the effect of these different groups based on the relationship between say

x and y. Say if you have gender over here what you can do is you can test these ah this and association among males and also among females. Now, if you have only two subgroup analysis it is fine, but if you have multiple subgroup analysis what happens is the chance of getting a significant result because you have just simply performed a subgroup analysis it goes higher and higher ok. So, it is always recommended that you perform this kind of a subgroup analysis based on what you have mentioned in the protocol that is pre specified and also limit the number of subgroup analysis.

Remember if you have large enough sample size then you can go on with different subgroup analysis, but your if your sample size is not large enough then as much ah subgroup you create there is chance of ah you know spurious association. So, actually the effect that you get ah you think that it is the moderation of that particular variable, but in fact, it is simply this spurious association and no moderation actually exist. So, this is the caution regarding subgroup analysis. Now how do we analyze the moderation and mediation then to be free from all those fears of ah this kind of subgroup analysis. One of the more emerging techniques is structural equations model here you have variables 1 to the outcome say outcome 1 you have the in intermittent variables I 1, I 2, I 3.

Now what happens is I 1 may have an effect direct effect on this relationship you can assess this through the linked regression ah equations this becomes the moderation analysis what can happen is V 1 to I 2 then O 1. What happens here is you can test whether I 2 is mediating the effect from V 1 to O 1 or not also what you can do is you can test whether V I 2 is also moderating along with mediating the effect of V 1 to O 1 or not like this. So, here what happens is you have line diagrams and based on the line diagrams the software it actually calculates different regression equations. So, structural equations model is again ah very ah emerging thing and it gives you a good result without the fear of the different subgroup analysis and the p values becoming spurious significant. Now we come to the last part of this lecture just to give you a brief overview most of you already know that there are different quantitative analysis software these are the different softwares that we use the spaces SAS, Stata etcetera we you have to pay for using these softwares while for R and Python you have different codes they they are kind of open source and you do not have to pay anything.

But the important thing over here why do we have to mention this statistical software is when we are proposing our research we also have to mention which kind of software and which typical software are we going to use and you have to cite that appropriately. So, all of this should be mentioned in the proposal phase and it should not deviate. So, basically we have discussed a lot of thing in this lecture to summarize we understood that there are different good practices for observational and experimental research data analysis we have to first describe the data we understand the different outcome variables we understand what level of measurement the outcome variables are in and based on that we choose the different statistical tests and we clearly have to understand the outcome of the test and for that we also understood that odds ratio relative these are the different measurements that we prefer to use with the categorical outcomes. We also understood that the relative risk is a better measure than the

odds ratio our discussion also included that it is always recommended to avoid unplanned testing in your research although in observational study you can indulge in a bit with the with the unplanned tests to formalize or find out ah certain some more information, but for experimental research it is always not recommended. And finally, we discussed that the subgroup analysis you need to limit the number of subgroups and a better way of understanding modulation mediation may be through the structural equations models.

So, these are your learning resources. Thank you for your patient hearing. Goodbye.