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 51: Quantitative analytical methods (Part I)

Hello again and welcome back. So, we were discussing the different aspects in Health Promotion Research. And towards the end of this course in this week, week 11, we will be discussing the different analytical methods that we employ in the health promotion research. Today our topic of discussion is the quantitative analytical methods that we are going to use in our health promotion research. This is typically divided in three segments, today we will be covering the first part of it. And for today the first concept that we will be discussing is the recapitulation of the matrix of measurement.

Remember we described and discussed the matrix of measurement previously during when we started to realize how do we measure the variables, how do we measure the concepts and constructs like these. So, we will have a quick recapitulation of the matrix in measurement because depending on the matrix that we use we will use different types of analytical methods. Then we move on to the description of a data that we get from our research and then we briefly introduce you to the concept of bivariate and multivariate analytical methods. Now, the bivariate and multivariate analytical methods we will be discussing a bit more in detail in the subsequent lectures.

So, what about the matrix of measurement? If you remember this is the same table that I used during the previous lecture when we discussed the matrix of measurement. So, this is just for a quick recapitulation. So, we have four levels of measurement or as we usually call them the four levels of data, the nominal one, the ordinal one, the interval and the ratio. So, what is the nominal data? It is based on certain criteria you can name them, but you cannot really rank them as higher or lower and you cannot really put a value to them. For example, if you consider the different genders then you really cannot put a score or a value to them and you cannot rank them like higher or lower this is better than this, this is worse than this like this you cannot put a tag to it you just have a name that is why it is called a nominal data.

Next you have the ordinal data, for ordinal data you have the order of preference like this one ranks higher. For example, a higher degree of pain a much higher degree of pain is ranked above a higher degree of pain and which is again ranked above neutral to pain stimulus like this. Next you have the interval and ratio data. Now, interval and ratio data as we have discussed earlier as well they have a particular score and the score typically has a meaning. For example, in interval scale as we discussed that say we have a temperature of 70 degree Celsius and 50 degree Celsius although that and we have a temperature of 30 degree Celsius.

Now, 70 degree minus 50 degree is 20 degree Celsius and 20 degree Celsius minus 30 degree Celsius is again 20 degree Celsius sorry 50 minus 30. So, this means that the differences between the two measurements they are same. So, the interval remains same, but the interesting part is here the amount of heat that is actually there in that scale or the difference in heat in terms of 70 degree to 50 degree Celsius and 50 degree to 30 degree Celsius that is not equal. So, that brings us to the next question of ratio scale where we have a true 0 what actually means a 0 in the in the in the scale of measurement of temperature we consider the Kelvin scale the 0 Kelvin means its absolute loss of heat. So, we can consider it as a ratio scale.

Similarly, if you consider the example of number of people who are in a particular room now in that case also you are considering the count of the people. Now see the concept of count is again very interesting because usually when you count an individual you get the whole number. So, that means, this is a continuous data it goes on from 0 to infinite, but again you know what happens with count is you get it in a discrete way that means, you can have a count of 1 2 and 3 and you cannot have anything in between. Now in this case also if you have a 0 that means, no personnel in that particular room that means, it is a true 0 like this. So, that becomes a ratio scale and see in a ratio scale you do have the properties that you have with an interval scale that means, the differences from a higher value to a lower value may be equal to a similar kind of a gap right.

So, based on these metrics of measurement we now move on to description of our data. Now the data that we usually get from our health promotion research we first entered the data in an excel sheet or in a spreadsheet software as we typically call them in terms of different columns and rows like this is the typical look of a data. So, here you have your variables these are your variables this is the serial number column here you have your observations like 1 2 3 you know in certain circumstances where you have a cross sectional nature of inquiry. So, that means, you are observing only one person once so that means, your observations or the serial numbers like 1 2 3 they are the serial numbers for the individuals of the participants and also that is the same observation that you have. Now here you have variable 1 2 3 4 like this.

Now in this part our objective is to describe the data. So, this is a typical display of the data that you get when you enter it in the spreadsheet software. So, how do we describe the data? See as we can understand that if we have 1000 representations or if we have 1000 individual points that we usually call them as the data points say we have this is I mean in this graph say we have the height of the individual we are now dealing with typically the physically measurable constructs ok. So, if we have the height of the individual and we have the number of participants for example, participant 1 comes here participant 2 like this we have it. Now the height may be distributed in this way ok.

So, it may be a densely distributed height. So, the interesting part is if you see this graph now if you are asked to describe the data or you know say what do you actually gather from the data then you then you really cannot describe all of them like individual 1 is having this much individual 2 is having this much that is not the way that is why we use certain parameters like the central tendencies the spread or the dispersion we use these different techniques. Now there are certain other techniques as well like frequency distribution we will come to it gradually. See for the height data what happens with height here is it has a typical value and you can consider height as a ratio scale as well. So, for the interval and ratio scale data what you can get? You can get an average value this line is representing the average value because you have a lot of numbers over here and typically you can get an arithmetic mean or an average.

So, that is what we call as the mean a mean can be calculated for any distribution assessed using the ratio level or interval level data we discussed that because in ratio and interval level data you have a typical scores and that scores really has some meaning. Now consider this when you are doing any study or whenever you have already conducted a study you consider the say the income levels in terms of the socio demographic or socioeconomic scales as 1 2 3 4 5 consider this example and you scored them as 1 2 3 4 and 5. Now for this particular scale if you take the mean of certain of number of people and you get the mean to be 3.5 now you can yourself understand that this 3.5 really does not make any sense what can be the mean level of a of this kind of a data which is not really interval or ratio level because the socio economic scale 1 2 3 4 and 5 they has an order to it that means, it is an ordered data or an ordinal data.

Similarly for the nominal data as well if you if you arbitrarily assign some score to a name and you take the mean of the scores it does not really make any sense because you really cannot take the average from apples and oranges and mangoes right. So, that is the basic concept of employing the mean as the central tendency measurement. Another measurement is the median perhaps median is more commonly used measure of central tendency in terms of health promotion research because in health promotion research we have different constructs some of them are measured in a continuous scale in interval or ratio level measurement and mostly they are somewhat measured in terms of the ordinal level data or the nominal level data. Now, when the measurement is usually typically the ordinal level or the interval and ratio level for all of these you can utilize the concept of median. What is a median? It is the score occurring at the midpoint of a ranked distribution of scores.

So, here the important part is ranked distribution of scores. However you can provide a ranked distribution of scores you can typically get a median and that median makes sense. See if you have a Likert scale of responses from 1 to 5, 1 being the lowest and 5 is ranked as the highest and next you measure a different individuals based on that ordinal scale and you get the median value say 3. Now that means, the median response to the Likert scale in that given population is at 3 or typically that is a neutral point. Now if due to some reason you get the median value after calculating as 3.

5 now that gives you a dilemma because typically you have device the Likert scale to take discrete values like 1, 2, 3, 4 like this. So, what does 3.5 in this scenario means? This is not

mean we are not summing up everything and we are not dividing it by the number of people this is the median that means, we have ranked the individuals like this is the rank and we are taking the midpoint of that ranked distribution and as we can conceptually understand since we can rank the ordinal data. So, the midpoint is quite feasible and it really makes sense. So, what happens if we get a median of say 3.

5 that means, the median response or the midpoint is in between the response 3 and response 4. So, in that scenario remember do not interpret the median as the value itself ok. So, that is the caution that we have to practice when we are actually interpreting the ordinal data. Now these parts you will get different types of representations and different types of discussions if you go through different textbooks typically of biostatistics, but you know this is the basic concept when we are going to describe our health promotion research data. Another concept is called the mode.

Now what is a mode? Mode is the most frequently occurring data that means, say we have numbers like 5, 10, 10 and then 3. Now in this distribution of 4 different data points you have 10 occurring twice so that means, 10 will be the mode. So, the mode perhaps is usually not used that much as compared to median and we subsequently will come to the normal distribution and everything. So, you will get an get a hold of the idea of how mode, median and the mean the concepts they differ and how they change. Now we have understood the central tendency that means, what is the midpoint of the data or the middle part of the data.

So, this line we have understood then we have derived through mean and mostly median. Now we have to understand how diverse how dispersed the data is showing because based on the dispersion we can really describe the data because this is the data as you can see over here this is the whole data. So, we now have the median or the mean and we now need to understand the dispersion of the data or the spread of it. One way to represent the spread is the range that means, the highest value and the lowest value you get the range by from in between them. You also have dispersion, dispersion is basically the difference between this scores, this scores means the values that you have over here for example, the height and their difference from the mean that is the dispersion or sometimes called the mean deviation as well.

The standard deviation is a typical measure of spread or dispersion it is the average dispersion from the mean so that means, you get the deviations of individual scores from the mean and you take an average typically the root mean square average that we called ok. So, that is how first we define the central point and then we define the spread that is how we describe the data. Now this is typically used for the ordinal, for the interval and for the ratio level data. The question is what do we use for an ordinal data because for an ordinal data the typical dispersion and the range these concepts although range can be used for the ordinal data 5 minus 1 the range is 4 like this, but dispersion the standard deviation that does not really make any sense. So, whenever we are utilizing median for our ordinal data we can include the concept of interquartile range that means, the 75th percentile and 25th percentile what whatever lies in between that is the interquartile range.

Again this is somewhat similar to the concept of range, range means only the highest values and interquartile range means somewhere in between the median and the highest value and the median and the lowest value and the differences between them like this. So, what happens with the nominal data? For nominal data we usually use the frequency distribution. What is the frequency distribution? It enumerates the number of occurrences for each of the attribute of a nominal measure. See for nominal data we are utilizing frequency distribution. Again if like like for an age if we distribute the age group say we have taken data from individuals to 18 years say to 45 years.

Now we have distributed this data in terms of 18 to say 30 years and 31 to 45 years. See whenever we have grouped this data into two now this 18 to 30 years and 31 to 45 years they are as good as a nominal data. For them as we can utilize the concept of frequency distribution we can simply count how many people are in this age group and how many people are in this 31 to 45 years age group. Accordingly we can prepare a frequency distribution table. Usually an ordered list of you know scores for one variable that also can be utilized in the frequency distribution.

So, that means, your frequency distribution table you can utilize it for the nominal data and also for the ordinal data, but whenever you are describing your data. So, that is why be very sure what level of measurement is actually used for that data, what is there in your protocol will be coming to the protocol part in the next week, but remember based on your proposal and based on whatever you have actually proposed to do proposed to measure based on that you can describe your data. It is not like a nominal level of measurement should be measured through means or like that ok. So, the concept must be very clear and the description should be accordingly. Next comes is the issue of homogeneity and heterogeneity.

What do we mean by homogeneity? Homogeneity classifies a sample of people who share a great deal of similarity for a specified variable. For example, if you if you go to a community and you are simply surveying the self efficacy beliefs for example, how effective or how well the individuals I mean they believed in themselves that they can perform a certain behavior. For example, sustaining tobacco quitting. Now if you are serving this question among a different group of people, now you can get different kinds of responses some people will say yes I am very much confident, some people will say no I am not at all confident some. So, some people will have a higher level of self efficacy and some people will not really have any self efficacy.

Now in this example typically I have given you a dichotomous choice that means, two options yes I can and no I cannot. Now in this scenario you tend to get a cluster that we call a cluster because for yes you have certain responses and for no you also have certain responses. Now if you consider the self efficacy in a scale of say 1 to 5, 1 means no self efficacy at all and 5 means the highest level of self efficacy and 2, 3, 4 in between. What you can get is you can get this clustering distributed in different suppose some people will say 1, some will say again 2, 3, 4 like this and 5. Now if they are distributed almost similarly in all these data points then we call it a homogeneous data because this is representing a homogeneous distribution.

If the distribution in these data points are not really that much similar say we have a higher degree of responses in 5 and a very low response in 2 again a very high degree of response in 1 say a very high degree of response in 3 so that means, it becomes a heterogeneous data because the distribution is not similar to in respect to this variable the this particular variable calling self efficacy. So in a way the concept is if you if you divide the variable or if you categorize the variable in only a few categories then you tend to get a more homogeneous response, but if you if you go on dividing the categories into further smaller categories then the true nature of homogeneity may be appearing and in fact, the data may not at all appear as homogeneous it may become heterogeneous as a whole. Now normal distribution is the basic concept that we have to understand when we are going for any data analysis. So, all of these curves they are called a normal distribution curve because as you can understand they are all similar they are all bell shaped curve that we typically understand from the concept of normal distribution. Now the next question is why normal distribution is really important? Firstly because whenever we are going to employ the different hypothesis test or the different you know tests of association with our health promotion data if our data can be distributed is distributed normally we can go for the parametric tests and if they do not require a normal distribution or in other words if the data not distributed normally then we go for non parametric tests.

We do have certain statistical tests to understand whether the data are distributed normally or not you can very well perform these tests in the different statistical software's that we have right. Another important concept here is see this red line that means, this line now for this mu means the mean and the sigma square it is the standard deviation so, the or the variance. So, for the red line you have mean at 0 and the variance at 1. Now this is called a standard normal curve we typically represent this with the letter z. Why this is important because again for the different hypothesis test and the different statistical tests the z test is or comparison with the standard normal curve is again a very important technique right.

So, this is again typical statistical things, but the concepts are very much important for you to understand how we can go about with the with the data analysis for our health promotion data. So, the crux over here is to understand whether do we need the parametric tests or we need the non parametric tests, rest the statistical software that you are going to employ it can perform, but you have to decide whether you need the parametric test or the non parametric tests. Next comes a very important aspect or a very important question of p value. Now p value we often you I mean use loosely and p value is a is the thing that we as researchers seek for p value less than 0.05 that means, significant that means, I have done good work, but exactly what is p value and does this really mean a very important thing that if a variable is significant and is it really that much disheartening if p value is not significant or not less than 0.

05. So, these are the questions that will be discussing in this slide. So, what is a p value? P value is the probability basically this is the probability that is why we denote it as p under the assumption that no effect or no difference that means, the null hypothesis we are assuming is there then the probability we are calculating the probability of obtaining a result equal to or more extreme than what was actually observed. So, that means, it measures how likely it is that any observed difference between the groups is due to chance why this is important because if the observed differences are due to chance then we really cannot infer that the intervention was perhaps effective. Say we have given in promotion intervention to different individuals in two groups. Now, whenever we are calculating the p value by different statistical methods and we get that the p value is 0.

04. Now, we assume that the standard level of significance is 0.05 we will again come to the come to the interpretation of why 0.05 later on, but for now let us assume that 0.05 is the is the level. Now, what happens here is this means that the intervention the changes that we get due to the intervention the probability of getting that by chance is 0.

05 which is which is substantially less which is less than 0.05 that means, it is a significant thing. So, the p value it really measures what is the what is the probability of getting the differences by chance because we test the alternative hypothesis that means, the hypothesis of getting a difference against the null hypothesis that means, of getting no difference ok. So, p value gives you the idea of what is the probability of getting the difference by chance I am repeating again and again. So, that the concept of p value I mean you get to understand what is actually the concept of p value.

Now, see a statistically significant test result that is less than 0.05 means that the test or the null hypothesis is false or should be rejected that means, it is the probability of chance occurrence is very low that means, the alternate hypothesis is true. Again a p value that is greater than 0.05 that means, there is no effect that means, we are accepting the null ok. So, that is the basic essence of having a significant p value and having a statistically not significant p value.

Also remember that if you get as p value which is not statistically significant or even if it is a statistically significant thing you do not get to infer how well the two variables they are associated with each other. So, that is again an important consideration that you should consider when interpreting the p value and if the relationship between two variable or the association between two variable when tested you get a p value of greater than 0.05 that means, not statistically significant then also you can keep on trying to I mean improving the sample size and the other design factors that you have over in your study to understand whether the relationship is actually there or not. If with a sufficient power and a good sample size you get a p value which is not significant then only you can conclude that the relationship between these two variables it is not statistically significant at it is not there in your current study ok. So, that is how the whole interpretation of p value goes on.

Next our discussion topic is the errors that we get in testing a hypothesis so that means, we have a hypothesis of A equals to B that means, it is a null hypothesis and the alternative hypothesis is A is not equals to B we also have two different alternative hypothesis A is lesser than B and A is greater than B. Typically in these two situations we use a one tailed test that means, we get a normal distribution we have discussed that we compare everything on the statistic that we get from the statistical test in terms of the normal distribution here we have the two tails these are the tails and typically what we consider we consider a 95 percent that is the two standard deviation this is this part whatever lies outside we consider it as by chance. So, if anything is lying outside and and that means, the probability is less than 0.05 see because 95 percent plus 5 percent is 100 percent. So, these area they amount to the remaining 5 percent or the probability of 0.

05 to put it simply right and for so, for A less than B or A greater than B what do we do we usually take only one tail like this one or this one depending on the direction of hypothesis that we are testing. Now, consider we are testing these two hypothesis that A equals to B and A not equals to B. So, this is the null sorry this is the null and this is the alternative ok. See what happens here is you can have two kinds of error one is called the alpha error another one is called the beta error how what can happen is actually you have to accept the null hypothesis that means, the data that you have or the interpretation or the result of the statistical test that you have that has actually occurred by chance, but you are unable to conclude that as that has occurred by chance and you are concluding that no the alternative hypothesis is correct that means, there is difference when actually there is no difference and another scenario can happen that there is some difference, but you are not able to detect that difference. So, these are the two errors and this is what we will be discussing in alpha and beta error.

So, you have your type 1 error and type 2 error the the alpha error is called type 1 and beta called the type 2. It is the error of concluding that there is something some change difference or effect when in reality there is nothing as such so that means, the null hypothesis was true, but you have erroneously concluded that the alternative hypothesis is true. So, there is actually no difference, but you have concluded that there is some difference that is the alpha error. So, the opposite scenario it will be the beta error that means, the concluding that there is nothing when in reality there is something ok. So that means, when you reject the null hypothesis when it is actually true you consider the alpha error and you fail to when you fail to reject the null hypothesis when it is actually false that amounts to the beta error.

So, I hope the concept of this errors in hypothesis testing is clear because this is again important when you are going to you know do a power analysis and the sample size and everything. Next we will be showing you the different relationship between typically two variables that we can have. See in this case the variables they are the the values are increasing with each other that means, if we consider this as the x variable and this as the y variable the value of y is increasing with the increasing value of x. So, this is called a direct relationship see the graph is moving like this ok. Now what happens here, here see this is x and this is y now we are not going into the type of data that we have over here just consider them as the continuous data typical continuous data that we have here the value of y is decreasing as the value of x increases.

So, the direction is like this and this is called an inverse relationship. Typically there can be another type of relationship where partly you get a rise and then you get a fall. So, this is called a quadratic relationship after a certain point the the value of y typically decreases with the value of x or in fact, it can happen like this ok. So, that is the issue with the with the quadratic relationship. Now why these relationships are important because the relationship that we have between the two variables that we usually assess through the technique called the bivariate analysis bivariate two variables ok.

Usually commonly in our health promotion practice we shall be utilizing the t test which can be paired or unpaired we shall be utilizing chi squared test and to measure the strength of association we shall be utilizing the Pearson's correlation. So, typically from t test and chi squared test what do we get we get whether any association is there or not typically that is the major interpretation of these two test and then from Pearson correlation we not only get the the association that is the p value we also get the strength of association through the Pearson's correlation and that association you know typically is depicted through these diagrams. So, how this is increasing or how this is decreasing how much that is represented by a value called r that we that is usually the Pearson's correlation coefficient ok. So, this is the idea behind bivariate the relationship between two variables. The last concept of our discussion is the multivariate analysis technique seen bivariate we have two variables and we are assessing the relationship between two variables and in multivariate we typically have different variables say v 1 v 2 and v 3 more than 2.

Here what happens is you have v 3 say as the outcome variable that means, v 3 depends on v 1 and v 2 like y dependent on x here v 3 depends on both v 1 and v 2 here you cannot typically form this kind of a graph typically if you have three variables you you will be having a three dimensional representation of data and that is what we utilize in the multivariate analysis techniques in health promotion. Some of the common techniques that we will be using we will be using the linear regression where the outcome variable that is the v 3 in this example is continuous in nature logistic regression when it is dichotomous in nature that means, kind of a yes no situation or a male female situation like this. Count regression where the count variable that means, it is countable the number of illnesses or say the number of people who are entering a particular OPD that is the count regression. There are certain advanced techniques these are a few advanced techniques like structural equations model it utilizes the path analysis framework, multilevel model where you have the data nested in different different levels of of certain variables that is called a multilevel model all of these all of these higher level of analytical techniques and the basic analytical techniques they fall into the multilevel analysis strategies. So, in conclusion the quantitative analysis it usually starts with the description of data we discussed how to describe the data the descriptive analysis first the central tendencies then dispersion we have to understand also whether the whether the data should be we should represent in terms of frequency distribution tables or not.

We discussed the importance of normality we have to choose whether we go for the parametric test or the nonparametric test and we understood the different relation types of relationships perhaps between the variables like direct inverse or quadratic we discussed all those aspects and also discussed the different available bivariate and multivariate techniques in a in a very short way. We shall be discussing a bit more about the bivariate and multivariate techniques in the subsequent lectures, but for now that will be all for this lecture and I hope you will be going through these resources because these are the areas from where we have we have presented these examples and illustrations. Thank you very much.