Exploring Survey Data on Health Care

Prof. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture - 24
Composite Index

Welcome friends, once again to the NPTEL MOOC module on Exploring Health Survey Data. We are on the week explaining on survey data analysis discussing in this particular lecture on Composite Index. I am just giving you 1 minute background of composite index to start with.

In research papers we know that if we just give the variable as it is, it may not have any analytical base or it may not create interest for the readers to relate to the reality. And the in depth of the paper may not be well defined therefore, we need to define or develop certain variables which are composite of different variables.

In reality we have so many latent variables. So, many variables which are not directly discussed, there is no single quantification possible just by the name of that variable. Suppose I say empowerment; empowerment if you look at that variable, it does not have any value directly, it is consisting of so many indicators.

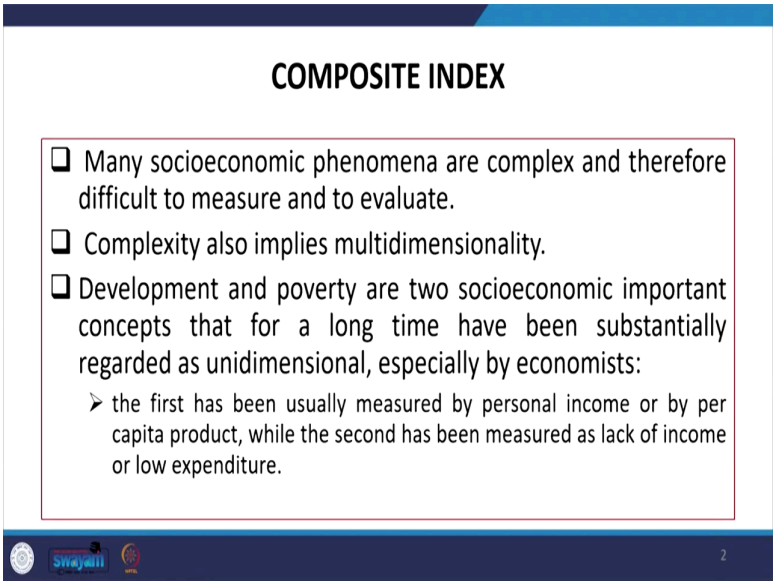
So, empowerment may be relating to decisions of the person or empowerment may be related to education of the person, empowerment may be related to the location of the person.

So, there are so many indicators that those are consisting of the index called empowerment, that is why empowerment is not directly observable. It is to be observed through different indicators and their weights. Therefore, we say this kind of variable called latent variable. And similarly, if we wanted to understand one variable called social capital; social capital has strong bearing on healthcare expenditure.

Social capital you may refer to the person, how the person is connected, how the social value the person is bearing. So, it is networking, it is indirect, the education, the community development etc. will be all composed together to define social capital. So, we are all going to discuss, we are gearing of to that extent and give you the direction of understanding composite index and also give you some practical background or practical steps to calculate and interpret composite index through PCA.

So, let us move on the first slide, I will discuss about how this composite index is developed, what are the background of this composite index?

(Refer Slide Time: 03:22)

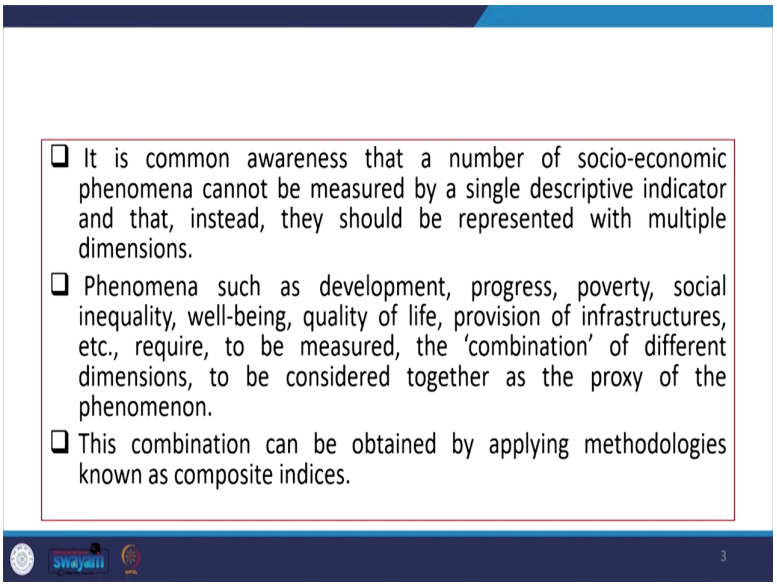


Many socioeconomic phenomena are complex and therefore difficult to measure and to evaluate. Complexity also implies multi dimensionality, development and poverty are two socioeconomically important concepts that for a long time have been substantially regarded as unidimensional, especially by the economist.

But in fact, it is not unidimensional it is composed of so many dimensions, when we develop and go by only one aspect like development and poverty it is in fact, multidimensional in nature.

So, the first, that is, development has been usually measured by personal income or by per capita conception or product. While the second that is poverty has been measured as a lack of income or low expenditure, that is why those might be be proxy, but not entirely discussing poverty. So, it you might have heard about multidimensional poverty index, that is what is presently discussed in most of the policy documents.

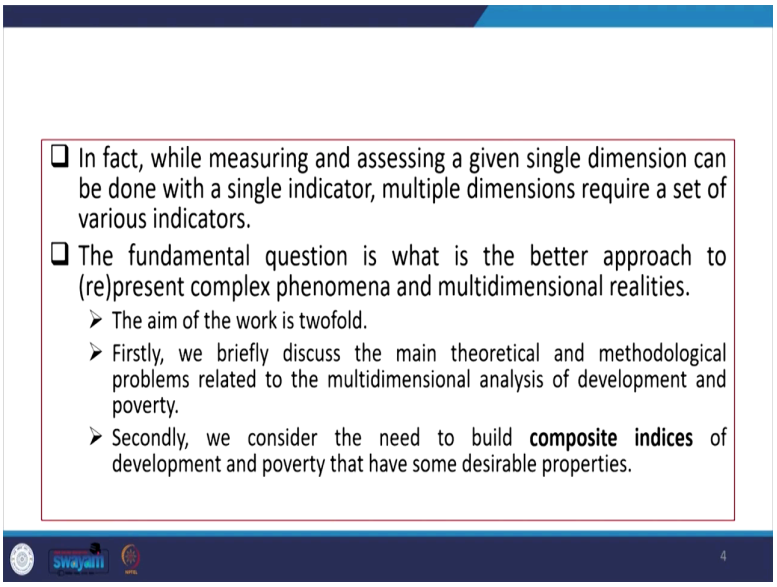
(Refer Slide Time: 04:33)



It is a common awareness that a number of socioeconomic phenomena cannot be measured by a single descriptive indicator as and that instead, they should be represented with multiple dimensions. Phenomena such as development, progress, poverty, social inequality, I have already mentioned about empowerment, social capital, quality of life, well-being, provision of infrastructure, etc. are called latent variable are not directly observed.

And this required combination of different dimensions and they consider together to define the word which we have just cited. These combinations can be obtained by applying methodologies known as composite indices.

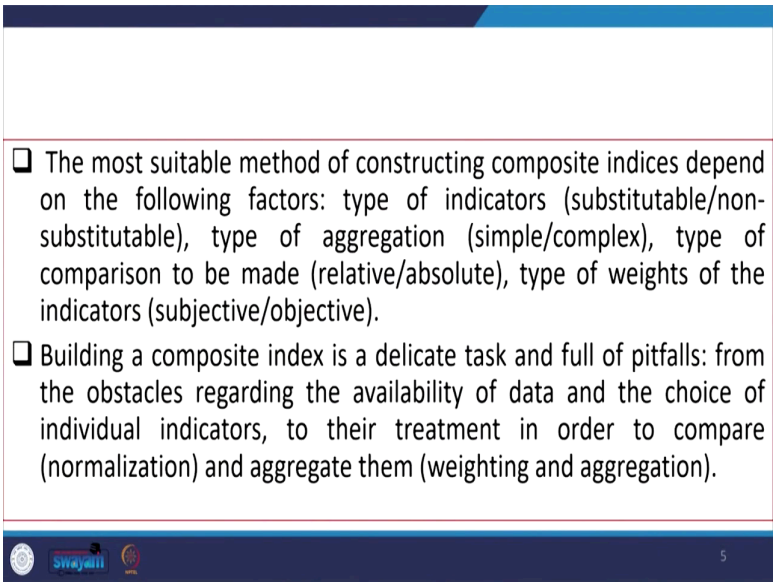
(Refer Slide Time: 05:21)



In fact, while measuring and assessing a given single dimension can be done with a single indicator, multiple dimensions require a set of various indicators. The fundamental question is what is the better approach to present or represent complex phenomena and multidimensional realities? The aim of the work is twofold; Firstly, we briefly discuss the main theoretical and methodological problems related to the multidimensional analysis of development and poverty.

Secondly, we consider the need to yield a composite index or composite indices of these two variables that have been already discussed and we will be including its desirable properties.

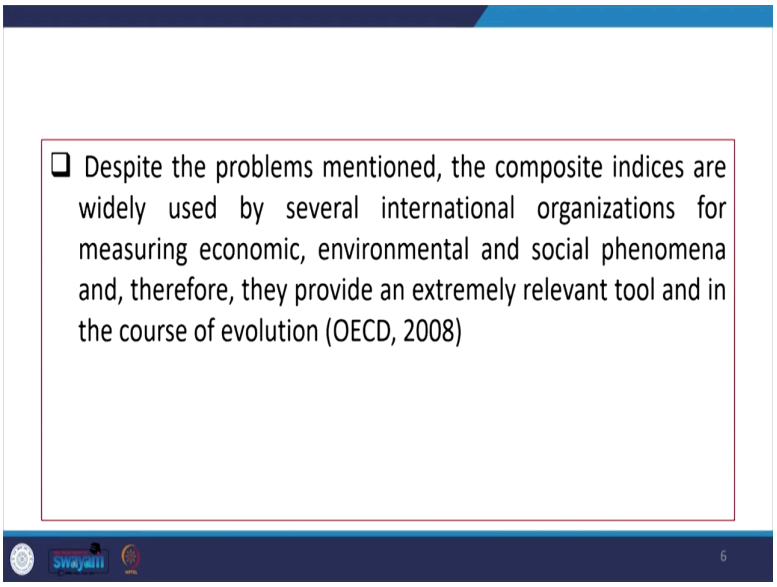
(Refer Slide Time: 06:18)



The most suitable method of constructing composite indices depend on the following factors: type of indicators, maybe substitutable or non-substitutable type of aggregation, simple and simplex, type of comparison to be made between relative and absolute, type of weights of the indicators maybe subjective or objective based.

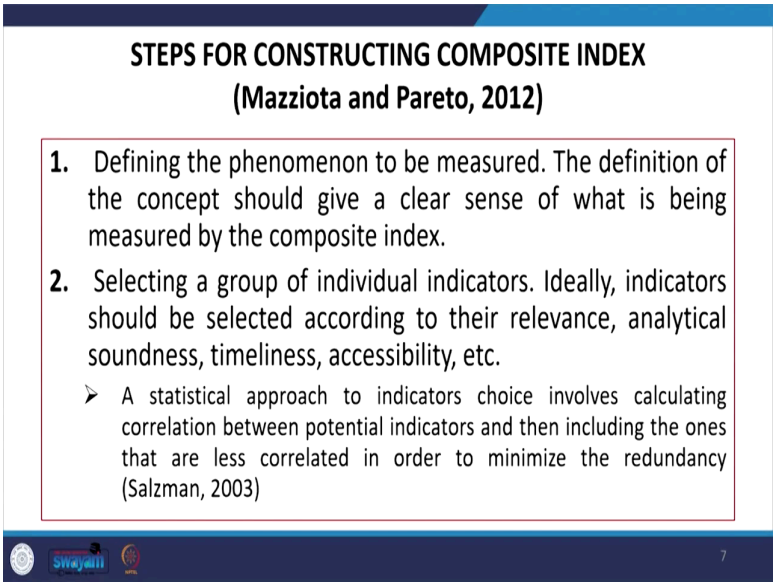
Building a composite index is a delicate task and full of pitfalls: from the obstacles regarding the availability of the data and the choice of individual indicators, to their treatment in order to compare or normalize with its normalizations and aggregate them with the help of weighting and aggregation.

(Refer Slide Time: 07:01)



Despite the problems as mentioned, the composite indices are widely used by several international organizations in measuring their environmental, economic, social context which, therefore, provide an extremely relevant tool in the course of evaluation site as mentioned by OECD in 2008's definition.

(Refer Slide Time: 07:24)

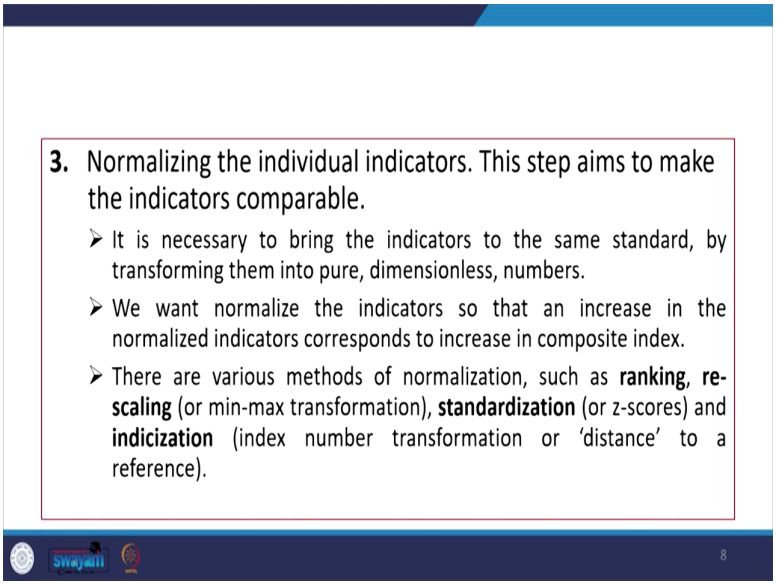


There are different steps involved to construct a composite index, a suggested by Mazziotta and Pareto in 2012 paper. First of all, defining the phenomenon that is to be measured, the definition of the concept should be a clear sense of what is being measured by the composite

index. Then selecting a group of individual indicators. Especially we need to take care of their relevance, analytical soundness, timeliness, and accessibility etc.

A statistical approach to indicators choice involved in calculating correlation between potential indicators and then including the ones that are less correlated in order to minimize the redundancy as suggested by Salzman 2003 paper.

(Refer Slide Time: 08:19)

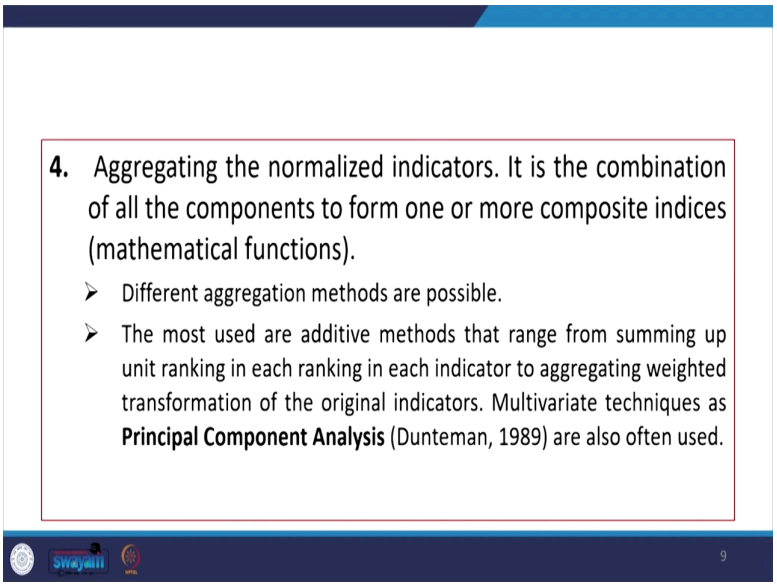


Normalizing the individual indicators are also important. The steps that aim at making the indicators comparable with the normalization, these are important. It is necessary to bring the indicators to the same standard, by transforming them into pure, dimensionless, and numbers. It has to be comparable; it has to be completely having certain scale, it should be unit free. Then only those can be made as a composite index.

We want to normalize indicators so that an increase in the normalized indicators correspond to increase in composite index. There are various methods of normalization such as ranking method, rescaling with the min-maximum transformation, then standardization with z-score (z-score minus the mean divided by standard deviation) and indicization.

Basically, some forms of index transformation are required like, distance to a reference, how far it is and then accordingly some forms of transformation is required for comparison.

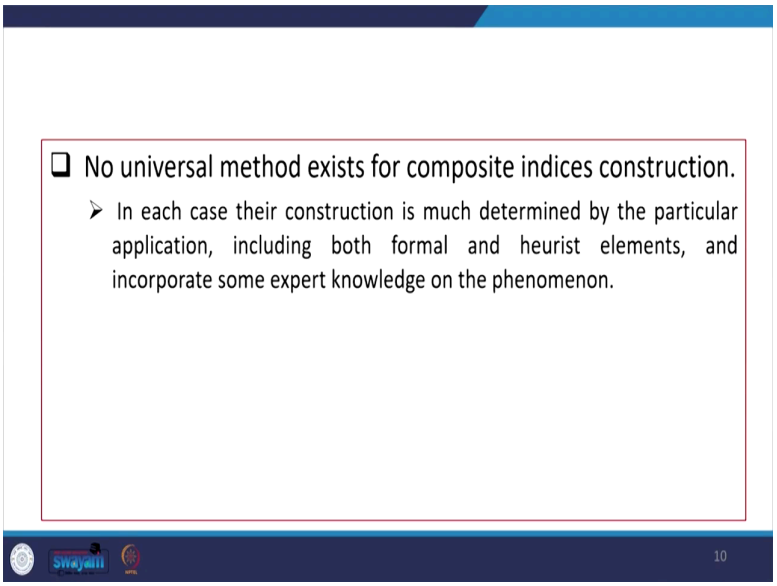
(Refer Slide Time: 09:38)



Aggregating the normalized indicators is the next step. It is the combination of all the components to form one or more composite indices, having its mathematical properties functions. Then different aggregation methods are possible.

The most used are additive methods that range from summing up unit ranking in each ranking of the indicators, to each of the indicators to aggregating weighted transformation of the original indicators. Multivariate techniques are called principal component analysis are often used, has also cited by this author.

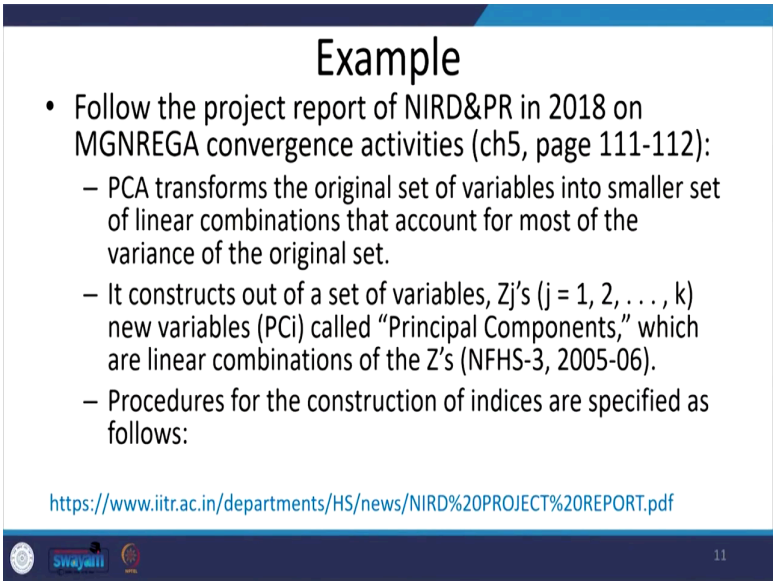
(Refer Slide Time: 10:24)



No universal method exists for composite indices construction. So, we should not just be confused that the indices or the composite indices is the best or the worst, should. You need not get yourself confused there is no such single universal method that is widely accepted by all, rather PCA to a large extent is widely used.

In each case their construction is much determined by the particular application, including both formal and heuristic elements, and incorporate some expert knowledge on the particular phenomena.

(Refer Slide Time: 11:04)



Now, I am citing one study of ours, one of our projects published and funded by National Institute of Rural Development and Panchayati Raj Institutions. Our team calculated and this is in the public domain, I have already given the link here. You can cite the page number 111 to 112, that is on MGNREGA convergence activities.

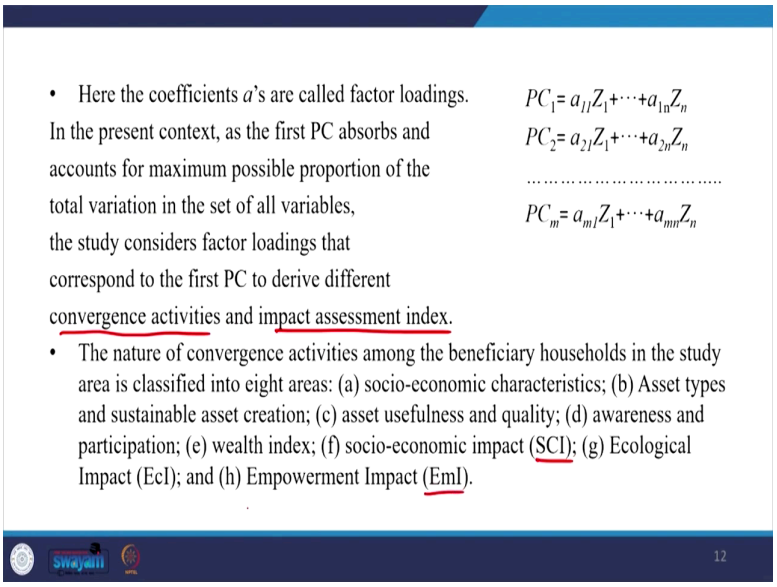
So, where we have discussed about PCA Principal Component Analysis? That PCA transforms the original set of variables into smaller set of linear combinations that account for most of the variants of the original set. If there are variants, already attached in the distribution of the variables. So, those are actually normalized and the most of the loading of the variables are considered in constructing a PCA index.

It constructs out a set of variables, like Zj's varies from 1 to k with new variables they are named as PCi called principal components, which are linear combinations of these Z,

availability the kth, availability of the j variables. You can also refer to NFHS-3 (2005-06) report which gives detailed discussion on PCA as well. Procedures for the construction of the indices are specified as below.

Here the coefficients, that is '***a'*** or '***α***' are called factor loadings.

(Refer Slide Time: 13:02)



So, accordingly the principal component scores are defined, in the present context as per the project report or project variable, the first PC that is the principal component absorbs and accounts for maximum possible proportions of the total variation in the set of all variables.

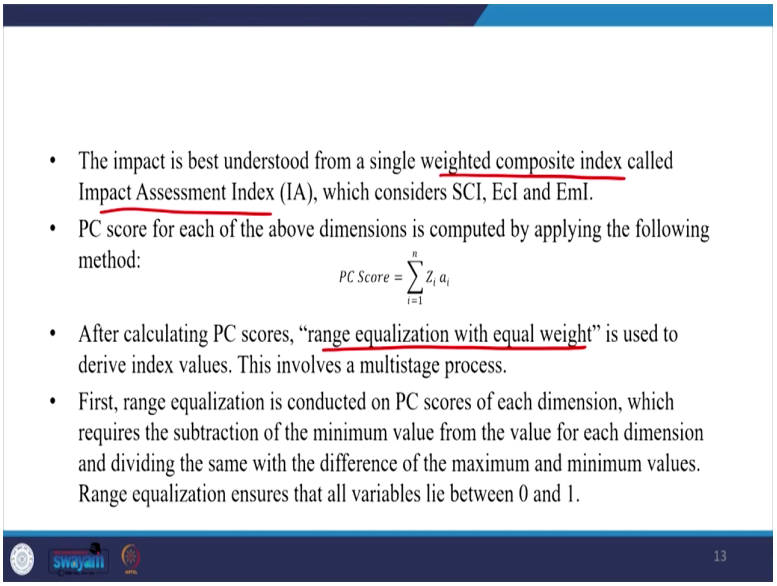
First of all, we need to do some factor analysis then, if there are so many factors if you do not have so many factors then directly you can consider those variables. And try to normalize the variables wherever there are some possibility of extreme values, make the distribution more normalized.

Then these first PC absorbs and accounts for maximum possible proportion of the total variance in the set of all variables. The study considers factor loadings that correspond to the first PC to derive different convergence activities and impact assessment index. We actually develop convergence activities index and impact assessment index, because the project was born on impact assessment of convergence activities in MGNREGA.

The nature of convergence activities among the beneficiary households in the study area is classified into eight areas. We identified that is, socioeconomic characteristics, asset types

and sustainable asset creations, asset usefulness and quality, awareness and participation, wealth index, socio economic impact, in short we have mentioned as SCI. Then ecological impact (EcI) and empowerment impact as (EmI) etc.

(Refer Slide Time: 14:37)



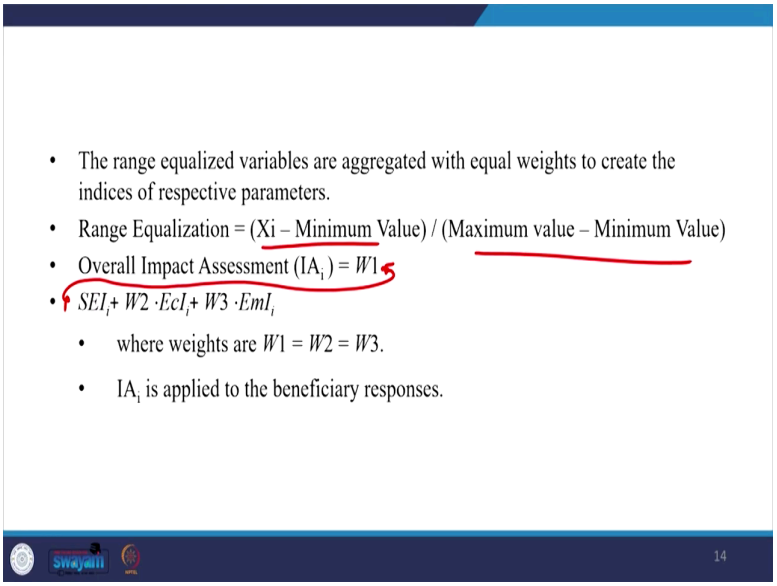
The impact is best understood from a single weighted composition index that is here as per the report is called Impact Assessment Index which considered these three important aspects SCI, EcI, and EmI. These are the things I have already mentioned, PC score for each of the above dimensions is computed by applying the following method. A summation of Zi ai their factor loadings, where, i vary from 1 to n.

After calculating PC scores, "range equalization with equal weight" is used to derive index values. That you need to also mark very carefully, range equalization is made because of having proper scaling of those variables, without having much deviation from the central mean.

And with equal weight has been considered to develop the index. This involves a multi- stage process. The steps involved are like, first the range equalization is conducted on this PC scores of each dimension, which requires the subtraction of minimum value from the value of each dimension and by dividing the same with the difference of the maximum and minimum values.

So, basically the equalization technique which we used to discuss. So, range equalization ensures that all variables lie between 0 and 1, that is why we said we need to make first of all, about those variables with range equalization.

(Refer Slide Time: 16:06)

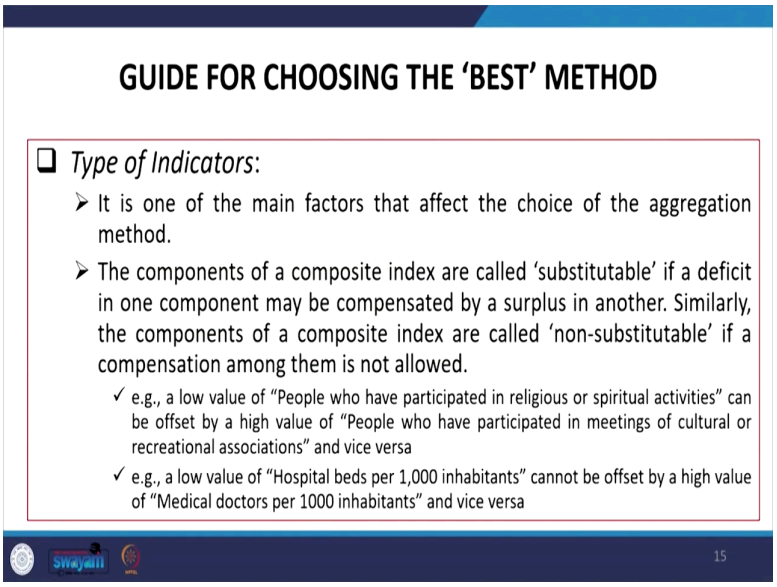


Then it is like, X minus minimum value divided by maximum value minus minimum value, this divided by this. So, this is what is called range equalization. So, range equalized variables are aggregated with equal weights to create the indices of respective parameters.

Overall impact assessment that is indicated as IAi is equal to SEI indicators that index plus its weight, basically this is multiplied here. It should have been taken this in continuation with this, this you can just read along with this. So, W1 it is weight of the 1st indicator, then weight W2 is the weight of the 2nd indicator and W3 is the weight of the 3rd indicator.

So, these weights are all equal i.e., equal weights have been considered. Then this impact assessment is applied to the beneficiary responses accordingly, how much impact assessment have been attached to each of the beneficiaries.

(Refer Slide Time: 17:13)



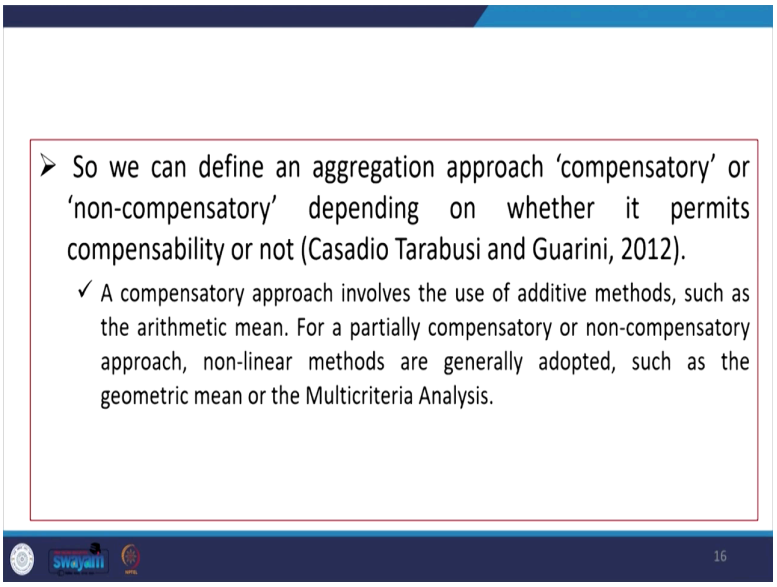
We will also give practical handouts here, for your better reference. So, the type of indicator which have been discussed will also reiterate. We are now guiding on choosing the best method on type of indicators, it is one of the main factors that affect the choice of the aggregation method.

The components of a composite index are called 'substitutable', if a deficit in one component may be compensated by a surplus in another.

Similarly, the components of a composite index are called 'non-substitutable' if a compensation among them is not allowed. For example, a low value of people who have participated in religious or spiritual activities can be offset by high value of people with have participated in meetings of cultural or recreational associations and vice versa.

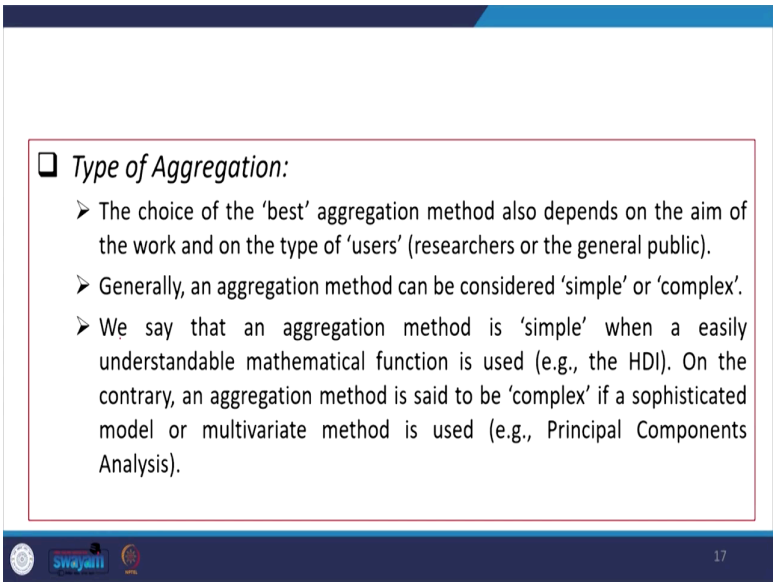
And another one is on a low value of hospital beds per 1000 inhabitants cannot be offset by high value of medical doctors for 1000 inhabitants and vice versa.

(Refer Slide Time: 18:28)



So, we can define an aggregation approach that is 'compensatory' or 'non-compensatory' depending on whether it permits compensability or not. A compensatory approach, in fact, involves the use of additive methods, such as the arithmetic mean. We usually use for a partially compensatory or non-compensatory approach, non-linear methods are generally adopted, such as the geometric mean or the multi criteria analysis. they are also adopted.

(Refer Slide Time: 19:00)

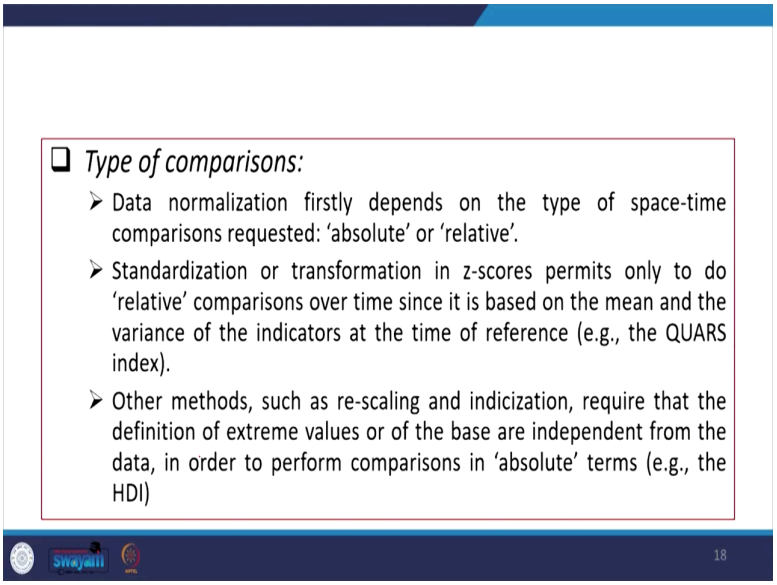


So, then type of aggregation is important, the choice of the 'best' aggregation method also depends on the aim of the work and the type of 'users'. So, type of users may be researchers

or the general public. Generally, an aggregation method can be considered as ‘simple’ or ‘complex’. We say that an aggregation method is simple when a easily understandable mathematical function is used. Let it be we are going to discuss about human development index.

On the contrary aggregation method is said to be complex, if a sophisticated model or multivariate method is used, that is principal component analysis. So, that actually includes multivariate methods based on factor loadings as well.

(Refer Slide Time: 19:47)

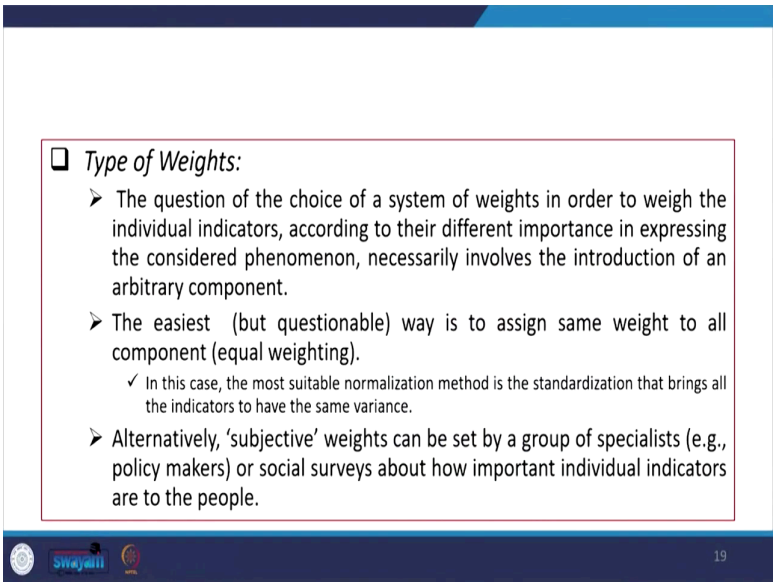


So, type of comparisons should be made; that is data normalization firstly depends on the type of space time comparisons between absolute or relative. Standardization or transformation z-score permits only to do ‘relative’ comparisons. This is what we have already discussed.

Then we can just read between the line, other methods such as re-scaling indices indicization indication, requires that the definition of extreme values or the base are independent from the data to perform comparison or in ‘absolute’ terms. That is what is HDI transfer and we will be also discussing this in detail.

What kind of weights are considered? The question of the choice of a system of weights to weigh the individual indicators, according to their different importance in expressing the considered phenomenon, necessarily involves the introduction of an arbitrary component.

(Refer Slide Time: 20:53)

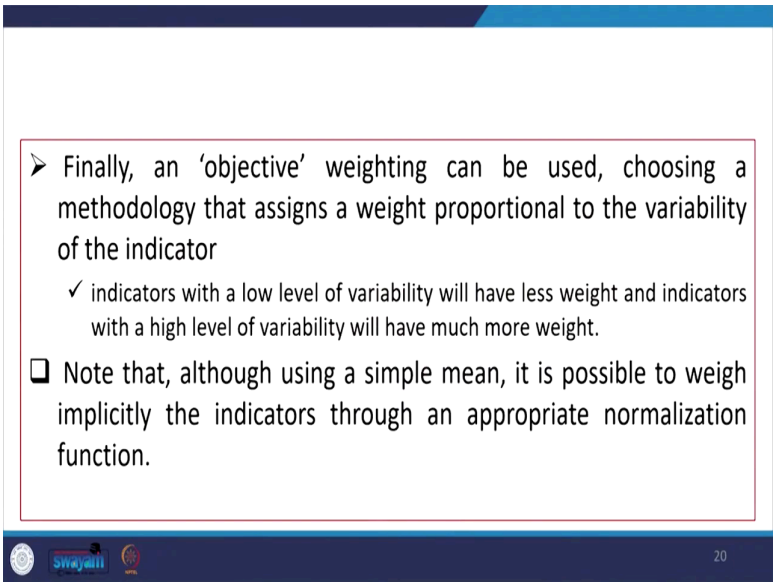


The easiest way is to assign same weight to all components that is equal weighting method. In this case the most suitable normalization method is the standardization, that brings all indicators to have the same variance.

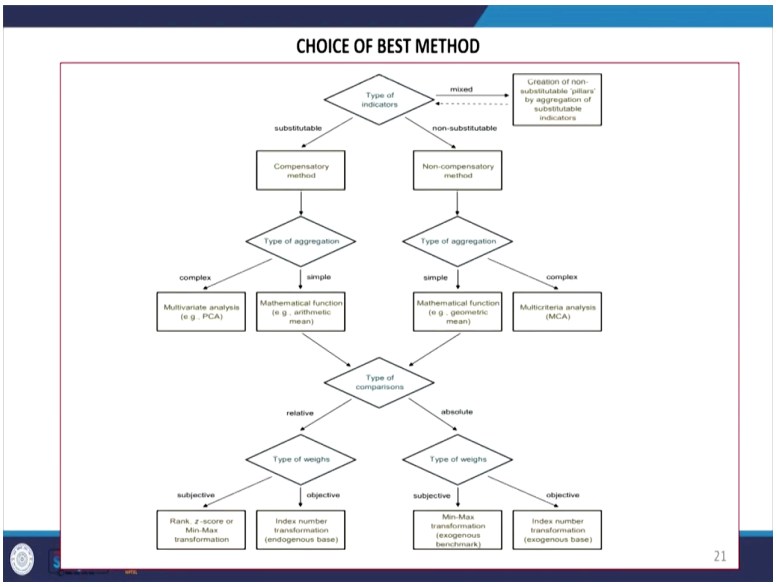
Alternatively subjective weights can also be given based on the loading of that particular group. And usually policymakers give and in social surveys about how important individuals indicators are to the people in specific indicators. Finally, an objective weighting can be used, choosing a methodology that assigns a weight proportional to the variability of the indicator.

You might have seen this variability weighting is torn in different large scale data set like NSS, NFHS. Indicators with a low level of variability will have less weight and indicators with a high level of variability will have much more weight.

(Refer Slide Time: 22:05)

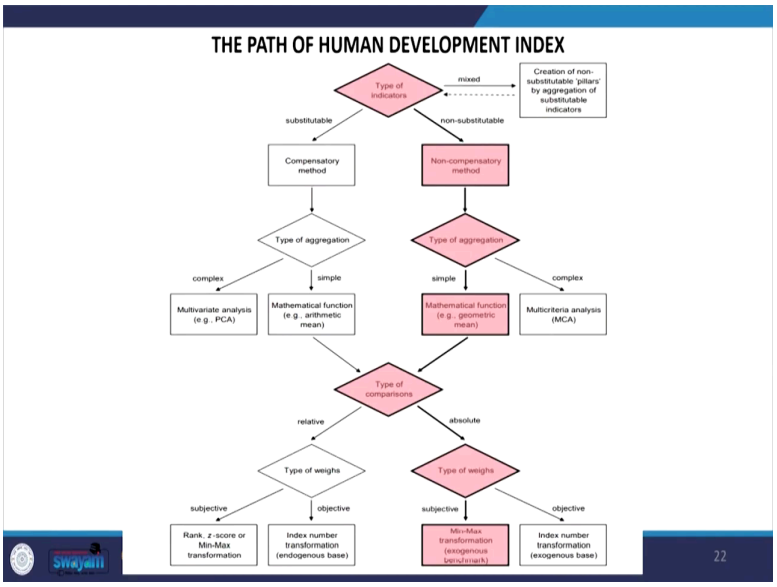


(Refer Slide Time: 22:10)



We need to note here that although we are using a simple mean, it is possible to weigh implicitly the indicators through an appropriate normalization function. Here I am citing the best method that is to be chosen, in developing index.

(Refer Slide Time: 22:21)



Next is on HDI, how they have followed the path. So, one we have already discussed based on types of indicators, there are various mixed methods involved. First of all, it is the type of indicators is based on substitutability versus non-substitutability, then this is based on substitutable indicators, they are called compensatory method or otherwise it is called non-compensatory method.

Then we should go for type of aggregation that we have already discuss and accordingly if it is complex then PCA is most used else simple mathematical arithmetic mean is applied. In case of non-compensatory method in their aggregation simple geometric mean is also perfectly fine. In complex case and it is called multi criteria analysis called MCA not as only PCA.

So, then we need to go for type of comparison, this is the first indicator, this is the second one, this is the third one. Comparison whether it is based on relative method or it is absolute weighting method. So, relative where we are supposed to discuss about the weights, whether it will be subjective or objective. Similarly in case of absolute method; subjective and objective method could be adopted.

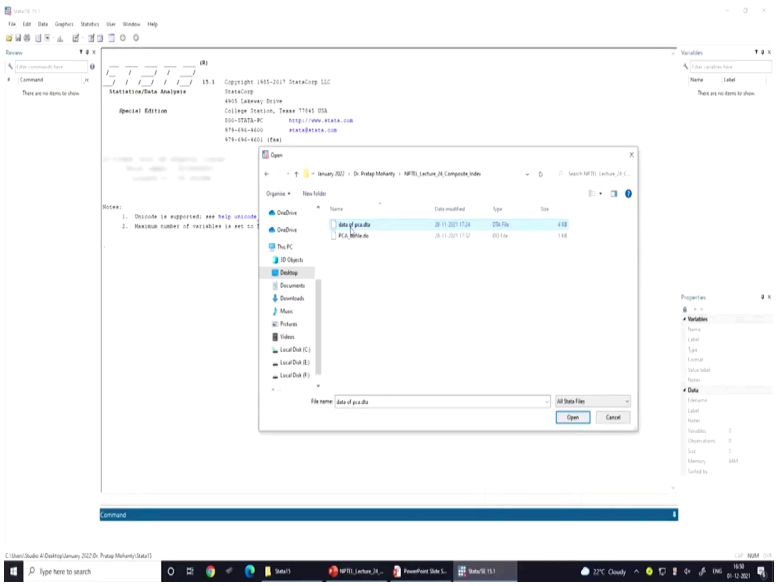
In case of relative comparison in the subjective weighting method are called rank, z-score, min-max transformation.

Similarly, in absolute one also min-max transformation is also used. Then in case of objective one index number transformation or endogenous based method are also used. Where in case of absolute one index number, extraneously determined weight could be given. This is how the path of human development index is followed. All those explanations are given how they have finally, followed and this is highlighted in the colour.

So, at the end after explaining all such things the min-max transformation method has been followed and the absolute weighting approach is also followed. And most importantly instead of making it complex they have used a simple method for aggregation, and they are non-compensatory method because the indicators are non-substitutable.

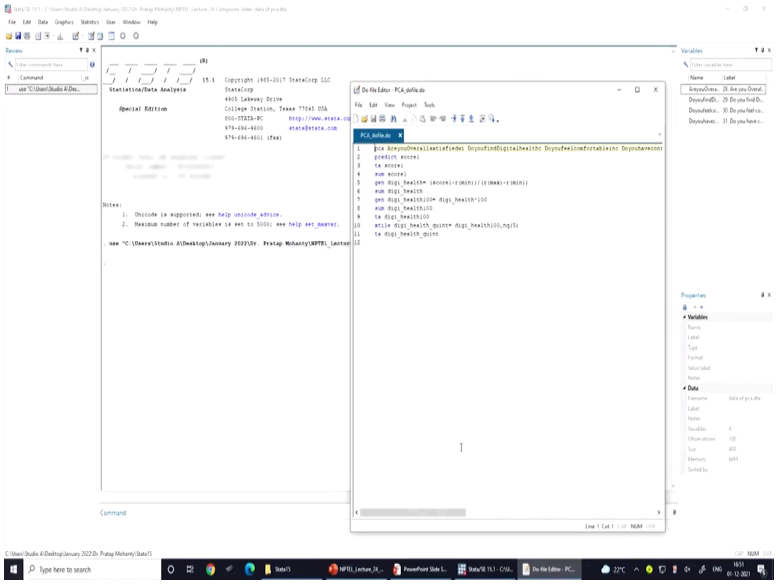
So, this is what the weighting or the composite index has been developed in HDI. Now, we are going to have some practical sessions of estimating PCA and how this could be interpreted. I will then come to discuss about some differences you can make after understanding this particular information. So, we are opening one practical sample data on the screen through Stata.

(Refer Slide Time: 25:33)



The Stata, the sample Stata subscription you might have all downloaded for 1 week, you can just save it and then operate with that sample version.

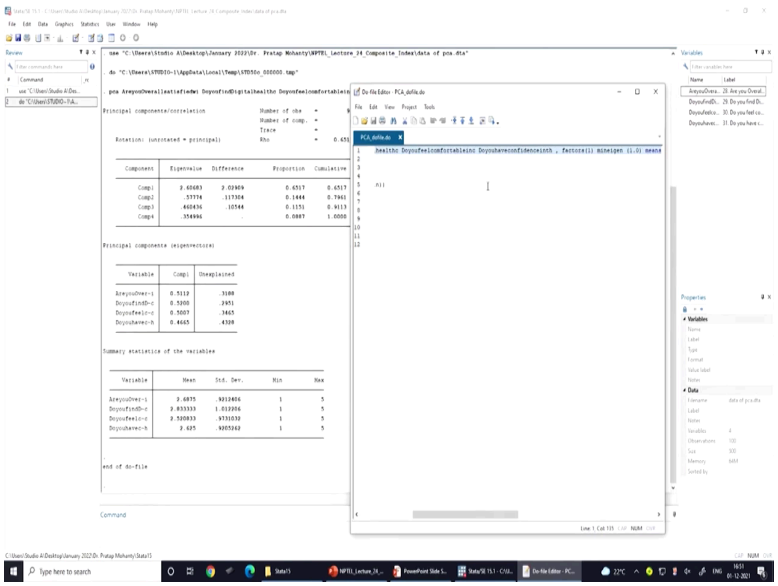
(Refer Slide Time: 25:47)



So, the data we have opened, the data we have taken from a field survey, from a primary field survey, maybe you can say for you it is arbitrary, for us it is the genuine one.

So, now, we are going to discuss about it, we will open a do file to understand how the index is developed. We will be including those 4 variables, first you need to actually define this PCA right.

(Refer Slide Time: 26:22)

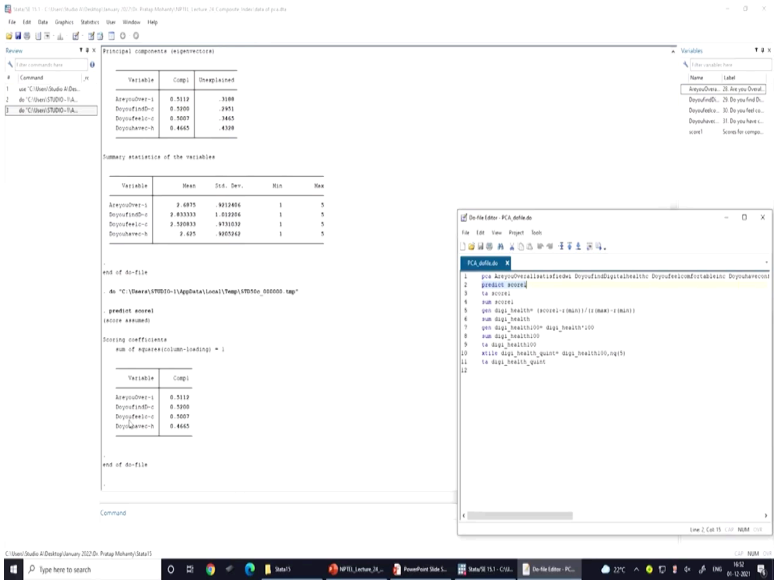


So, you need to define these 4 variables, how we have defined let me just mention you. This PCA is going to give you information about its factor loadings, Eigen values usually with higher Eigen values that components are considered to be having highest variability in the model.

Whichever the Eigen values is exceeding more than 1, they are considered to be having highest loading and there are principal components derived based on those Eigen values, each Eigen values are a kind of loading a kind of weight.

So, then what we will do, we will go for after defining all those things, the system is going to read its scores, predicted scores based on those factor loading.

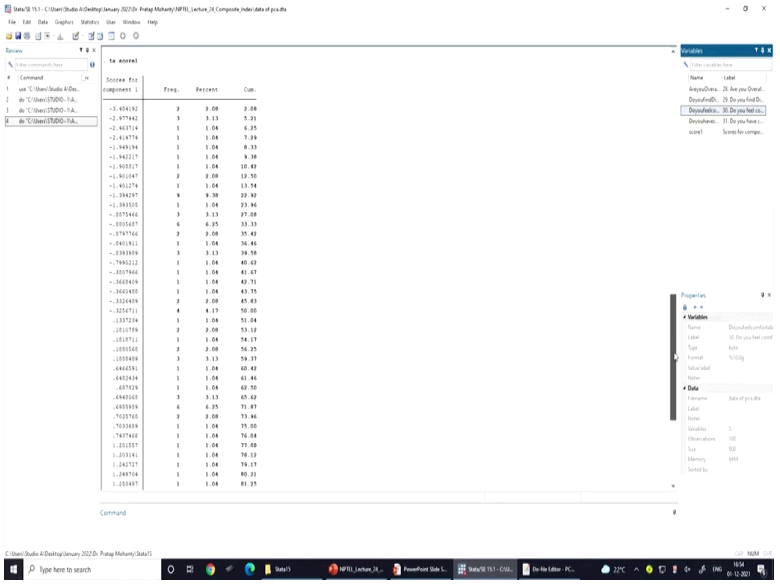
(Refer Slide Time: 27:17)



So, we will now operate the predicted score on your screen. So, these are the predicted score, these predicted scores like you can compare this principal component with the eigen values. Eigen vectors based on the eigen vectors, this one the simple, the simple principal components have been considered as the principal scores.

Then what we will do? We will in fact, discuss about we try to tabulate those scores once.

(Refer Slide Time: 27:50)

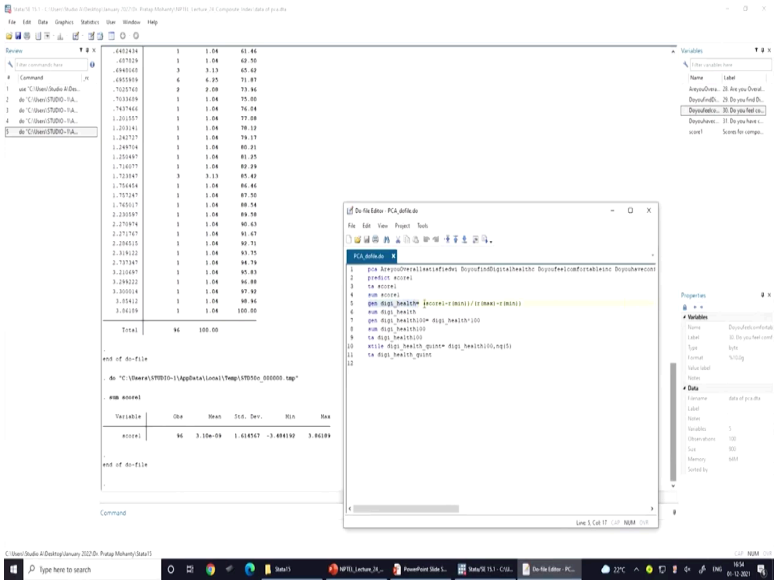


How these scores look like, starting from the end and we have the observations in our sample study of 100 observations. We are just mentioning this, what is this all about you need not understand, what are those variables like we try to understand the consumers perceptions, the individual perceptions about healthcare.

So, are you overly satisfied, just one variable I am just trying to make you understand. I am just taking tab each of the variables are here as is mentioned, you can just have a look. Do you have confidence in medical medicine prescribed through online prescription?

Here it is as; do you find Digital healthcare services as acceptable the way it is getting placed. Are you overly satisfied with the quality of services being provided. So, now we are calculating the scores after discussing all those things.

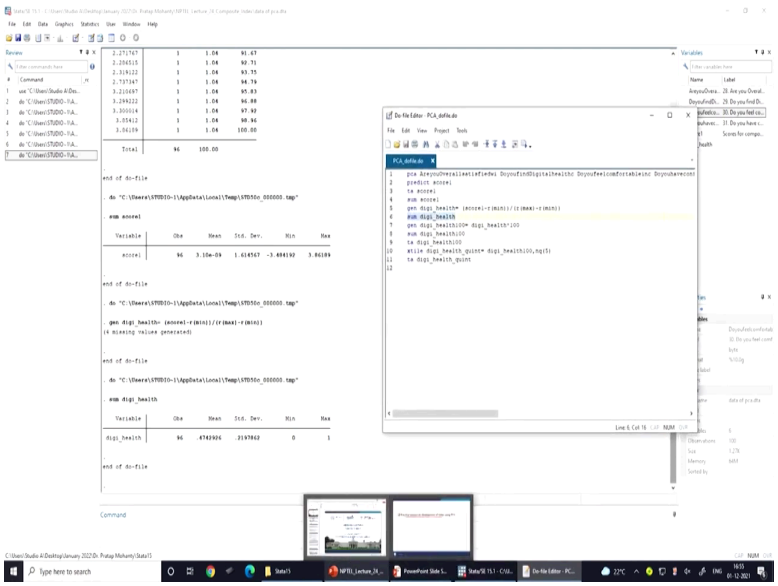
(Refer Slide Time: 28:55)



Here I am now going to summing up those scores. The summary statistics of those scores are calculated here. So, summary this gives us the mean values, number of final valid observations, then its minimum, maximum values all are here.

Then we will go for explaining as I already told you then next step is to define the normalizing the score, range equalization techniques should have obtained. Here we have applied the min-max strategy technique, that is score minus minimum divided by maximum minus minimum.

(Refer Slide Time: 29:36)

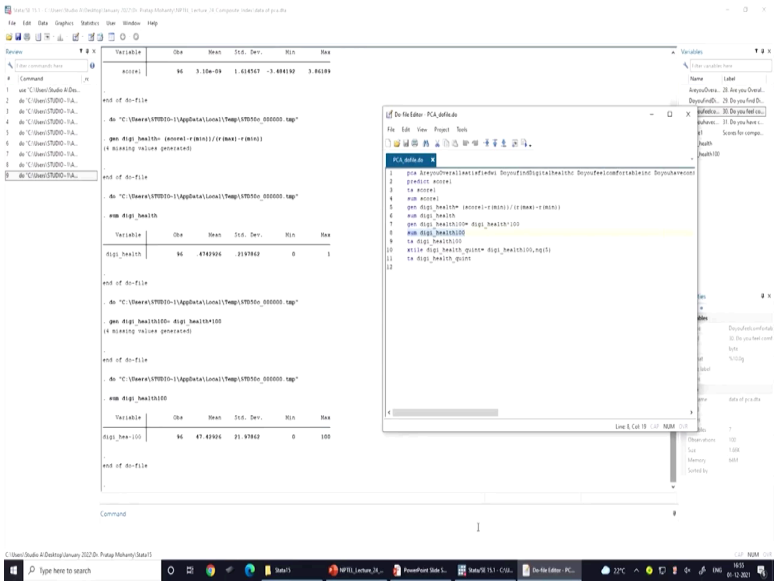


So, this is the one we are operating. So, 4 missing values are there, so 4 were not generated, rest 96; based on the 96 valid observations the normalization data has been defined. Now, again we are going to summarize it. Summarize the new variable we have defined as digital health or in short called digi health.

In the summary you can just see the mean expenditure, then mean score not expenditure mean score related to perceptions, their standard deviation, their minimum, maximum values.

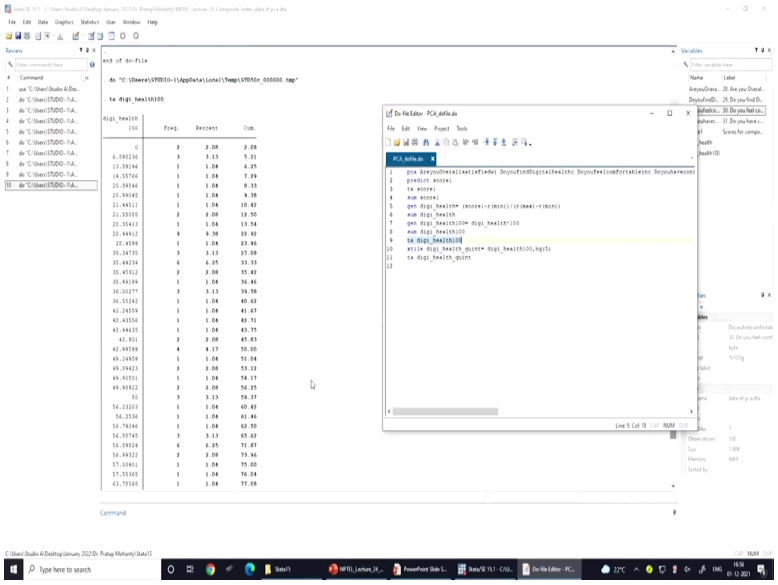
Now, all the values have been considered between the range of 0 and 1. Now, we are going to define or scaling it out of 100, instead to define an index. So, we are multiplying with 100. So, we have defined this, once again we can also summarize. Now, we can see 100 will be your maximum score and 0 will be your minimum score.

(Refer Slide Time: 30:36)



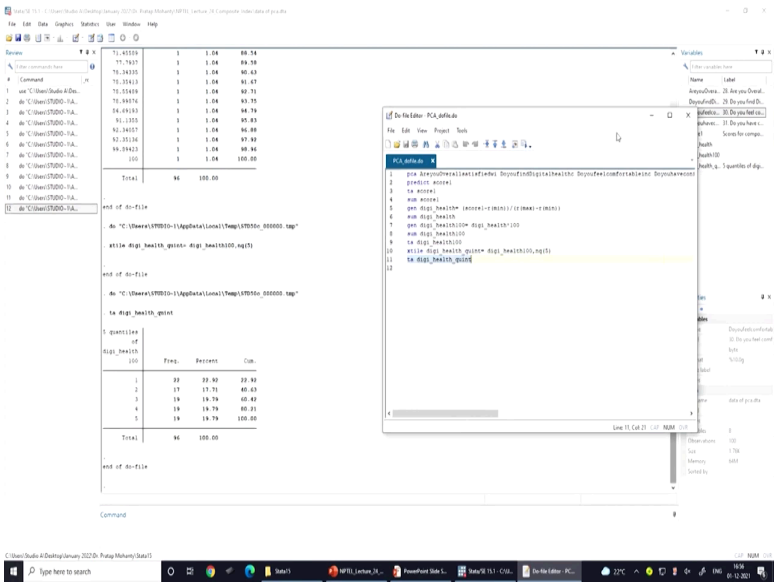
So, a minimum and maximum, since we have multiplied 100.

(Refer Slide Time: 30:47)



Now, we are going to tabulate that once again, we just want to see how it looks like. Initially it was in fraction, now it is in some in absolute values. So, we are now making the index. So, the command here is called ***xtile*** with a new variable name we have defined, we are defining in quintiles, this index in 5 categories. Index is a number fraction of number, but we are categorizing into 5 quintiles.

(Refer Slide Time: 31:17)



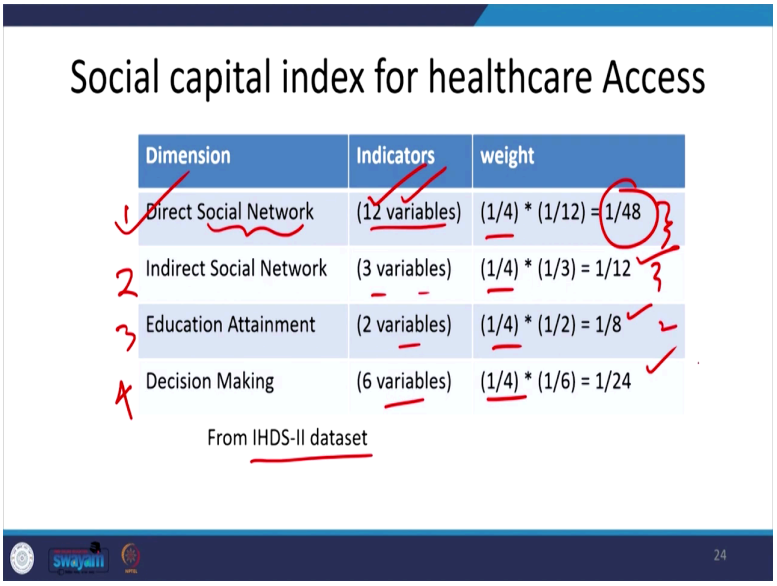
So, nq 5 is given, if you want to keep it is three categories then nq 3 has to be mentioned. So, at this moment it is 5, we are just running it and you can just see and how it looks like? We are going to tabulate and get the frequency distribution of it. So, here it is on your screen.

So these 5 indicators are on your screen, we have defined the composite index called digi health in quintiles. There are 5 categories and each categories how much their percentage of it, how much in each category you can easily understand.

If you are still a lit bit confused you can refer to NFHS 3 round and how wealth index is defined, wealth index they are defined into 5 quintiles. So, in which quintile whether a lower section of the society having highest frequency or higher whether the middle incomes are having higher frequency or not you need to get yourself confident and they have applied the PCA techniques.

So, after saying so I am once again going to open up the discussion to one of the sophisticated modelling we usually do, which is even better than the simple PCA.

(Refer Slide Time: 32:41)



Here I am just clarifying it using the IHDS data, IHDS second round data of 2011-12. Information available in that round is on social capital information, there are so many indicators of social capital. We are discussing in our paper on social capital index for healthcare access. How social capital is impacting healthcare access, like if a person is

located in rural area that person may not individually satiated with the kind of health care is being provided.

The person might get huge information about the availability of the health care facilities in their locality through their gram panchayat, through their village leader through their educated persons, maybe through their teacher, school teacher. They get to know about Asha workers and their implementation, they get to know about vaccinations during COVID-19, whether this is available or not. So, healthcare access is well modulated or well developed or understood penetrated to the last mile through social capital.

So, we have developed a social capital index based on 4 important indicators as available there, you can also modify there is no stand alone index here I am saying there are possible direct social network, indirect social networks, educational attainment, decision making.

Direct social networks the like persons who have directly know the person may be directly involve as a teacher maybe directly as a politician or member in Gram Sabha etc. attending Gram Sabha. Indirect social network where they have certain acquaintances with those top people.

I am not mentioning the exact variables you can go through an IHDS 2, there are 12 variables given in this direction, though there have been written as direct or indirect we have defined it. I am not clarifying all the details at this moment, if you have difficulties we will address that at our live session or in our chat box.

There are 3 variables here, 2 variables and 6 variables. We know that we should have given equal weightage. So, since in total we have 1, 2, 3, 4 indicators. So, one-fourth is our actual weight, one-four times 12 indicators are there. So, 1 upon 12.

So, the weight of each of the indicators under direct social network is of 1 of 48. Similarly, in the second case each of one is having 1 of 12th, then 1 of 8th and 1 of 24th. It seems as if you there are differential weight, but differential weight by each of the indicators, but equal weight by the broad indicators right.

So, that is what you can interpret, you can convert that variable with adding or multiplying this one 1 by 48 against that one, then finally, you run this PCA technique that we have just discussed.

That will be giving you the social capital index and will be very robust so far as your model is concerned, for avoiding multi quality issues, for getting better r-square values. your model interpretation is going to be much better and then usually accepted in different journals. These are all details for at this moment, we look forward to your participation in the next class.

Thank you.