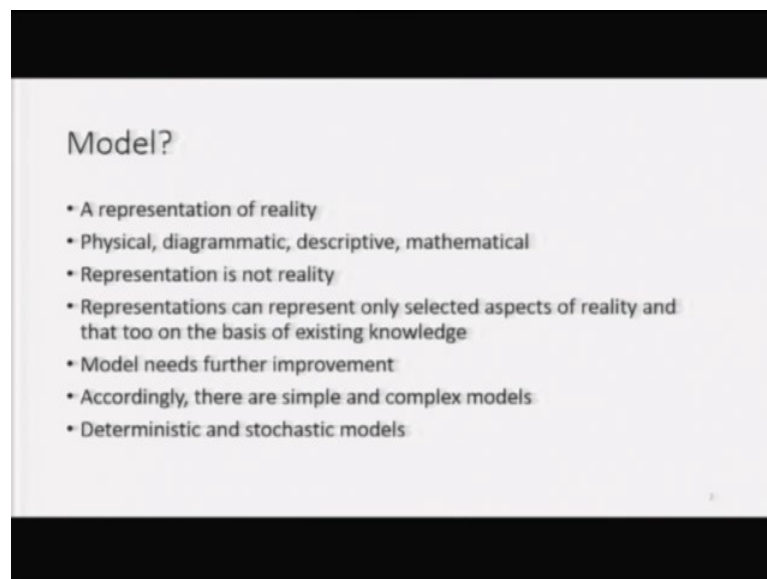


**Population Studies**  
**Prof. Arun Kumar Sharma**  
**Department of Humanities and Social Sciences**  
**Indian Institute of Technology, Kanpur**

**Lecture – 15**  
**Statistical Techniques in Population Studies - I**

Dear students, now two lectures are devoted to Statistical Techniques in Population Studies. I know that these days a lot of students from social sciences background are going for population studies and demography and they do not have the right background in mathematics and statistics. So, my presentation will be simple and I would focus more on what are the statistical techniques or what kind of statistical techniques are used in population studies and what kind of purposes do they serve. So, the main purpose of applying mathematical and statistical techniques is to build models.

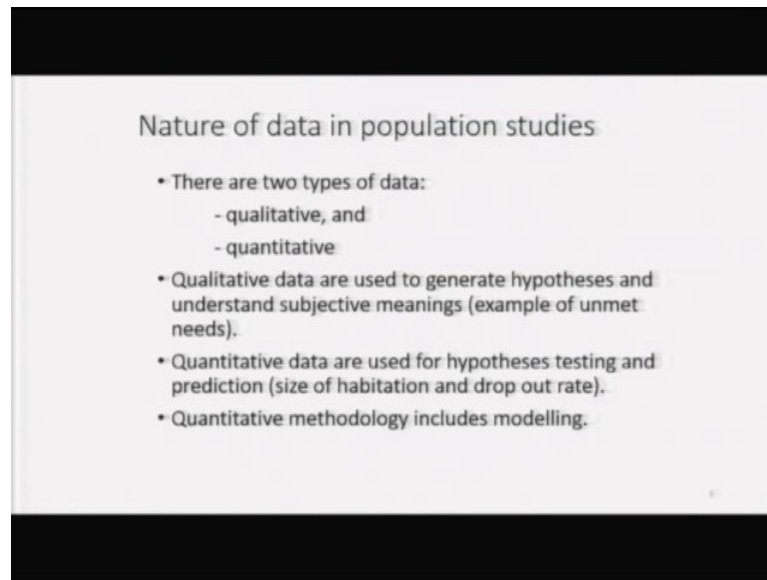
(Refer Slide Time: 01:14)



Models are a representation of reality, but not the reality. You choose certain aspects of reality and express them in mathematical or algebraic forms and that becomes a model. Models can also be physical diagrammatic, descriptive, mathematical. And these models or representations of reality represent only selected aspects of reality and that too on the basis of existing knowledge. All models need further improvement as more data come as more understanding of mathematics develops and as our expectations from models develop then models are improved.

So, accordingly there are simple models and there are complex models, there are also deterministic models and stochastic models. Deterministic models are those which deal with certain kind of variables while stochastic models are those which deal with probability events.

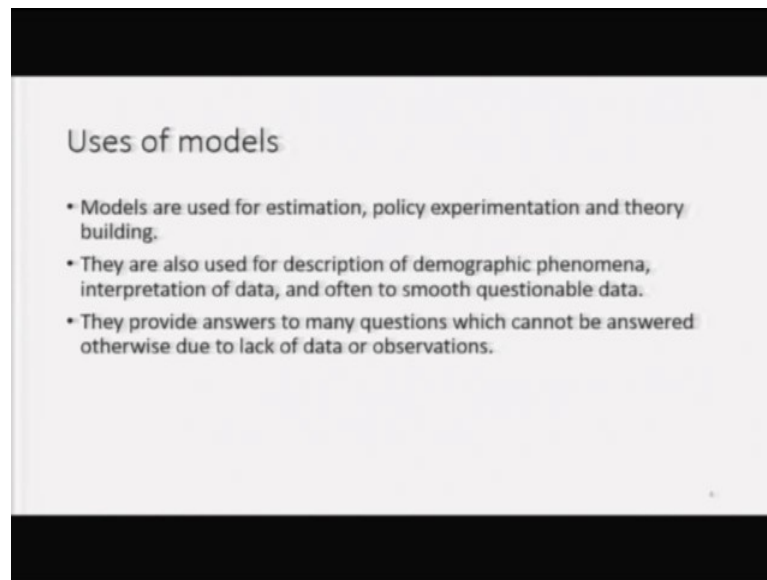
(Refer Slide Time: 02:32)



There are two types of data in population studies: qualitative data and quantitative data. Qualitative data come from ethnographic studies or from surveys when the variables are qualitative. They are used to generate hypotheses and understand subjective meanings in Max Weber sense.

For example, example of unmet needs that we discussed last time. Quantitative data are used for hypotheses testing and prediction size of habitation, dropout rate etcetera and this quantitative methodology includes modeling.

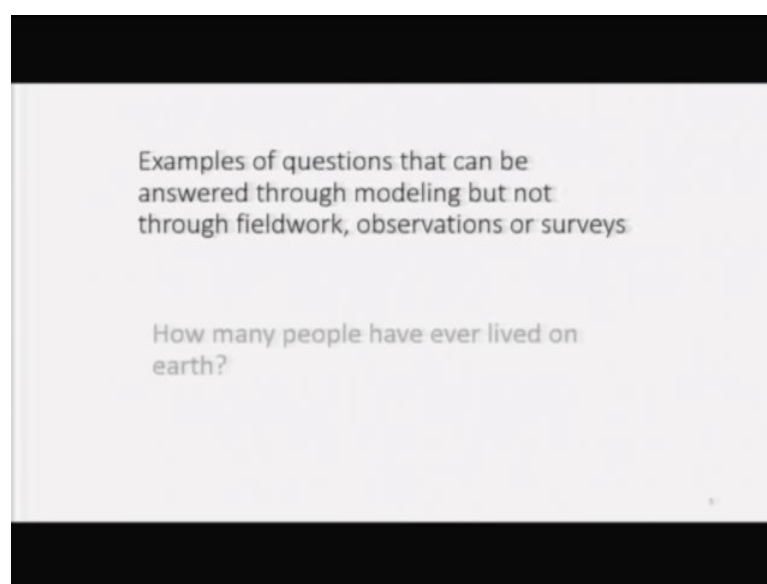
(Refer Slide Time: 03:25)



There are several uses of models. They are used for estimation, population policy experimentation and theory building. They are also used for description of demographic phenomena, interpretation of data and often to smooth out errors in data of questionable nature or erroneous data, data with known errors.

Models provide answers to many questions which cannot be answered otherwise due to lack of data or observations or because experimentation is not possible.

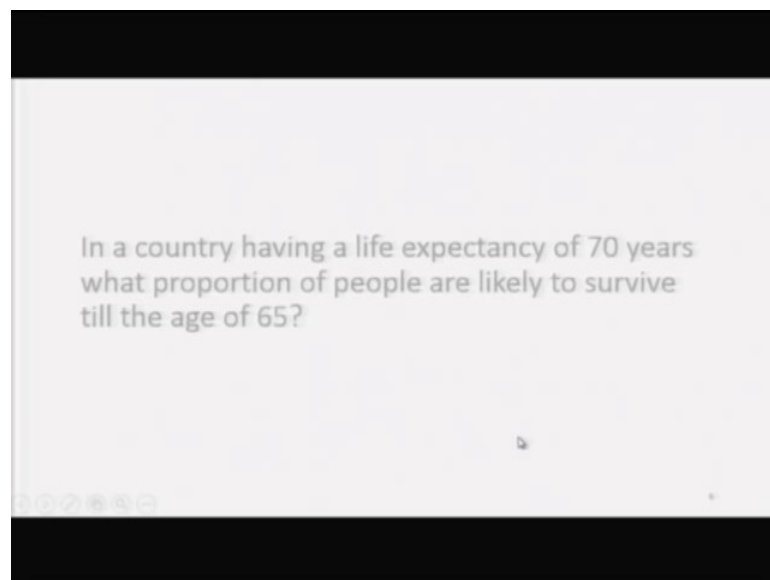
(Refer Slide Time: 04:16)



Examples of questions that can be answered through modeling, but not through fieldwork, observation or surveys; how many people have ever lived on earth. Now we do not have historical record of how many people have ever lived on earth, no survey can be conducted to estimate this number and no ethnographic work or any other source of data can answer this question, but models can. And some simple techniques developed by Nathan Keyfitz and other demographers have answered this question.

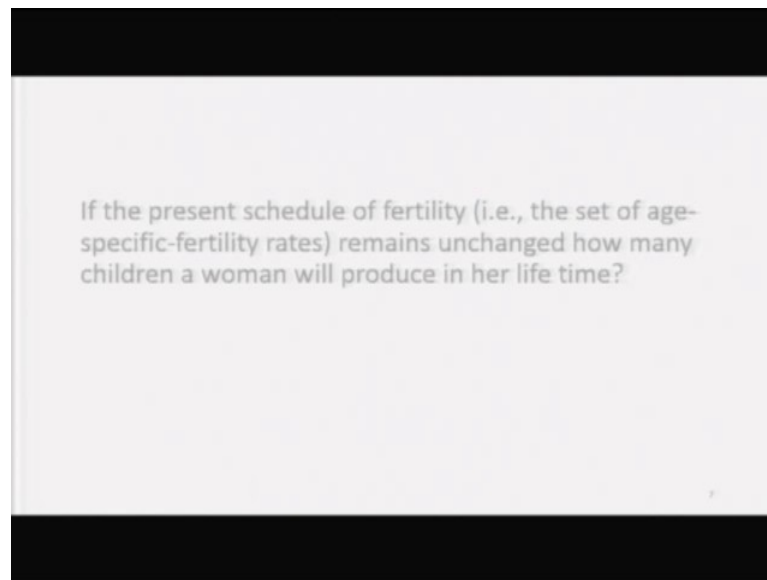
I think the answer would be that something close to 100 billion people have ever lived on this planet earth and the present 7 billion population of the world 7.7 to be more exact is 7.7 percent of all those who have ever lived on the earth.

(Refer Slide Time: 05:33)



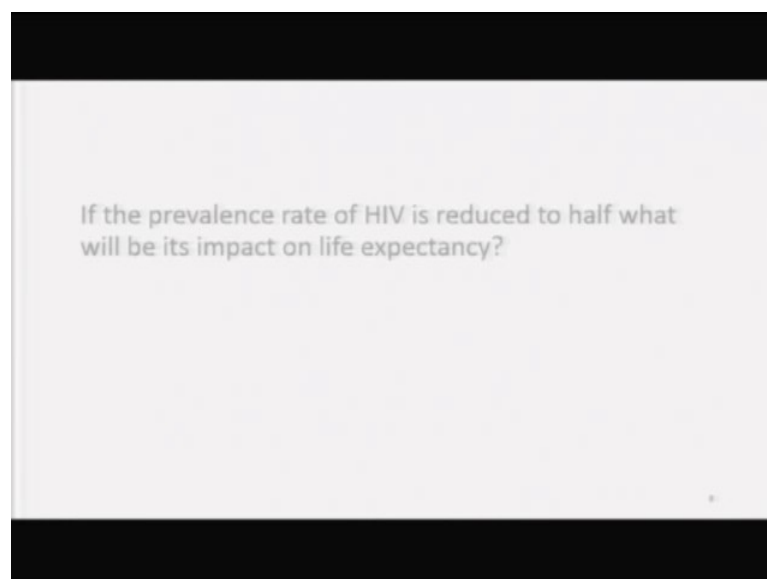
Another question which can be answered by building models is that in a country having a life expectancy of 70 years India is in that situation, what proportion of people are likely to survive till the age of 65 or those who are already of a 65, how many more years of life they can expect to live?

(Refer Slide Time: 05:52)



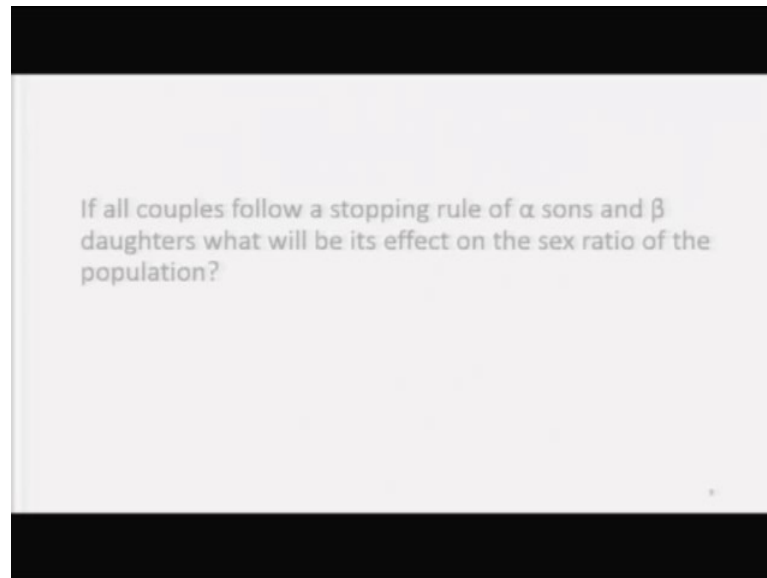
Another question if the present schedule of fertility by schedule; I mean the set of age specific fertility rates remains unchanged, you have age specific fertility rates of India for the year 2018 and imagine that the same rates operate for several decades, how many children a woman will produce in her lifetime? We have discussed total fertility rate and by adding age specific fertility rates, we arrive at total fertility rate and this is also a kind of model.

(Refer Slide Time: 06:49)



If the prevalence rate of HIV is reduced to half, what will be its impact on life expectancy? And when I was discussing life tables, I said that such questions can be answered by building multiple decrement life tables. Here we will need life tables according to cause of death and by using mathematical model and by eliminating HIV by 50 percent, we can recalculate life expectancy in that population.

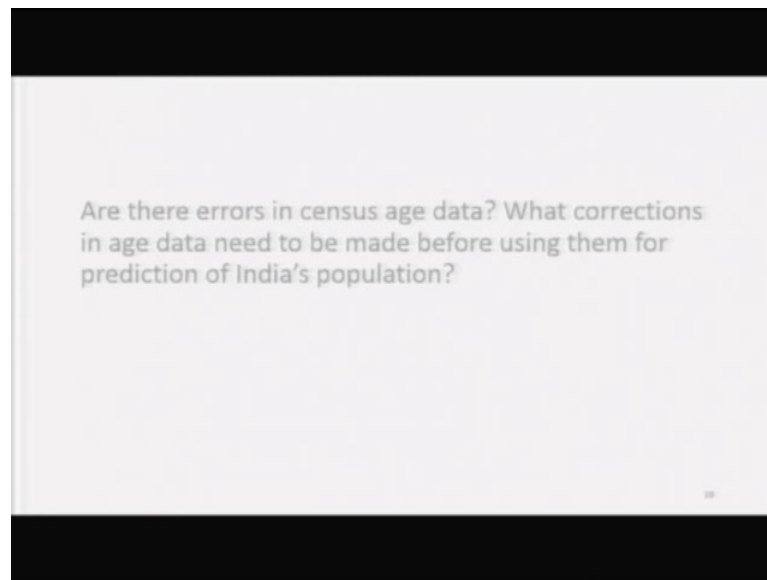
(Refer Slide Time: 07:24)



Another interesting question that can be answered by using model is that if all couples follow a stopping rule of alpha sons and beta daughters, what will be its effect on the sex ratio of the population? In 50s and 60s, the norm was or the stopping rule was two sons and one daughter couple. All couples or most couples wanted two sons and one daughter and the same norm continued for several years. Nowadays in urban area there are many couples who want only one child irrespective of sex. So, under these stopping rules, we can calculate how many children will be born on the average. It will not be alpha plus beta because all the couples will not be so, lucky that in first alpha plus beta children, they have alpha sons and beta daughters.

Some are still waiting for sons and some are still waiting for daughters and therefore, the average number of children will be more than alpha plus beta and that can be estimated.

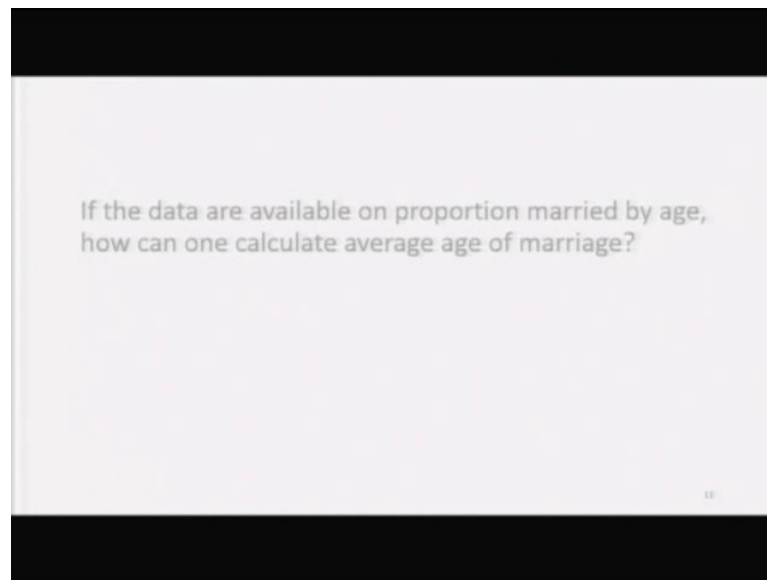
(Refer Slide Time: 08:39)



Another question is, are there errors in census age data? How can we know whether they are errors in census age data? We have a model of age distribution. Usually for correcting errors in age data we made use of stable population theory and a stable population provides age data for different values of life expectancy or death rate and growth rate of population. You can prepare age data for a particular stable population and compare your empirical data with the data of stable population.

And that will this comparison will show that at certain ages, your population empirical population may be underestimated or overestimated. So, you can correct your age data before using them for prediction of India's population.

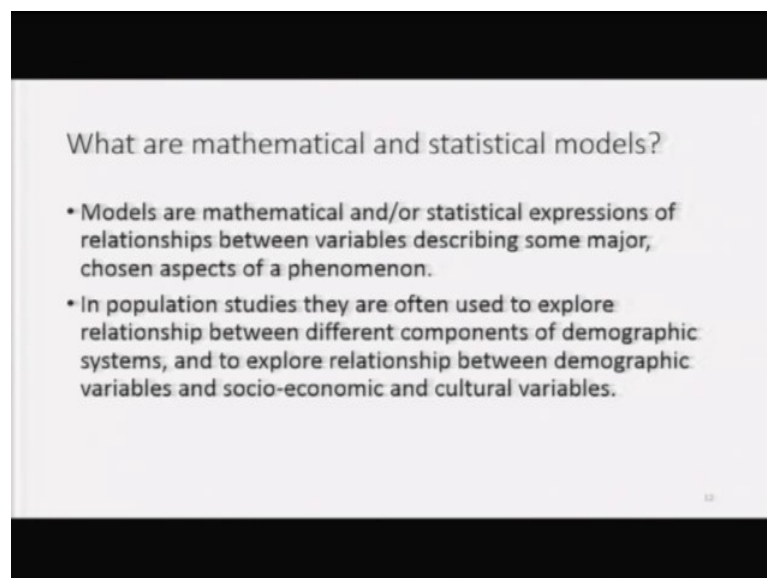
(Refer Slide Time: 09:46)



Another interesting question that can be answered by building models is that if the data are available on proportion married by age and census data provides such information for different ages or for different age groups, you know what proportion are unmarried, married, widowed, divorced, separated.

How can we calculate average age of marriage? And simulate age of marriage is one such way of calculating average age of marriage in that population. Now in general what are mathematical and statistical models?

(Refer Slide Time: 10:28)

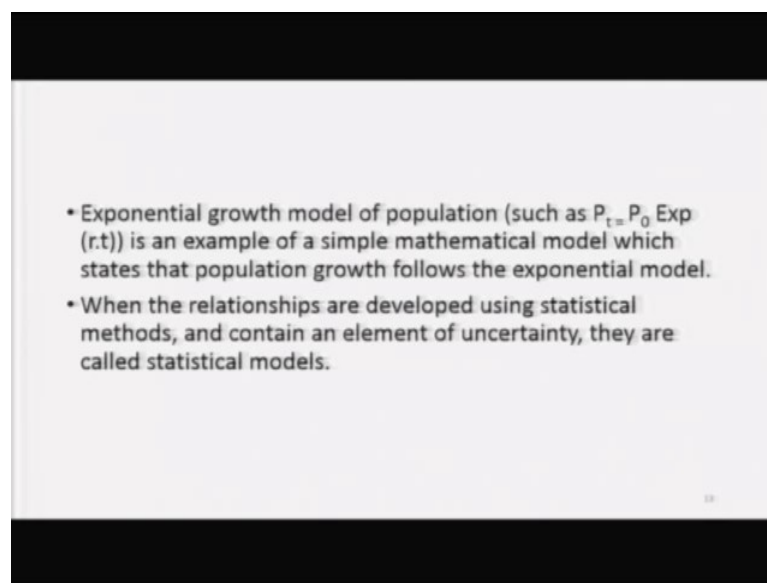




Models are mathematical and or statistical expressions of relationships between variables describing some major chosen aspects of a phenomenon. So, even a regression line is a model, exponential growth rate of population describes a model and life table is also a model, actually life table is called stationary model of population.

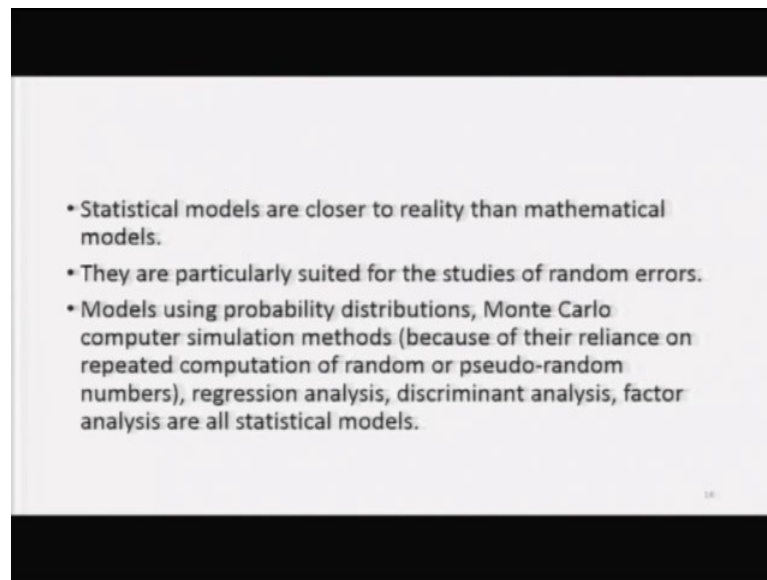
In population studies, they are often used to explore a relationship between different components of demographic systems and to explore relationship between demographic variables on the one hand and socio economic and cultural variables on the other.

(Refer Slide Time: 11:24)



Exponential growth model of a population such as  $P_t = P_0 e^{r.t}$  is an example of a simple mathematical model which states that population growth follows the exponential model. When the relationships are developed using statistical methods and contain an element of uncertainty probabilistic models, they are called a statistical model. So, mathematical models are analytical and statistical models are stochastic or wherever there is an element of probability or chance or random measurements or random errors we use statistical models.

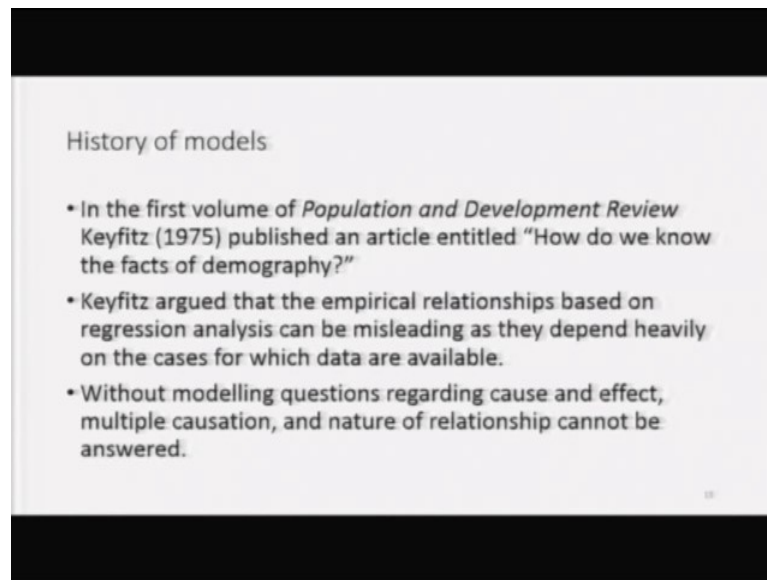
(Refer Slide Time: 12:12)



The statistical models are closer to reality than mathematical models. They are particularly suited for the studies of random errors, models using probability distributions, Monte Carlo computer simulation methods many demographic studies have been based on Monte Carlo computer simulation methods.

In these introductory lectures, we will not be able to go into these issues, but I am just indicating that if somebody has the right background and wants to know more on these things, then one can go through books on computer simulation. Because of their reliance on repeated computation of random or pseudo random numbers, then regression analysis, discriminant analysis, factor analysis and all statistical model; these days structural equation models or path analytical models are becoming more common in analyzing survey data.

(Refer Slide Time: 13:13)



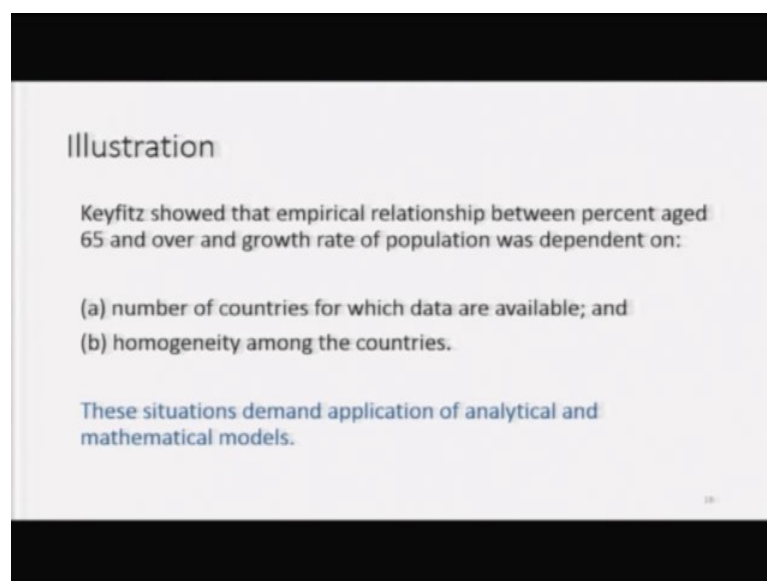
History of models

- In the first volume of *Population and Development Review* Keyfitz (1975) published an article entitled "How do we know the facts of demography?"
- Keyfitz argued that the empirical relationships based on regression analysis can be misleading as they depend heavily on the cases for which data are available.
- Without modelling questions regarding cause and effect, multiple causation, and nature of relationship cannot be answered.

13

A history of models in population studies: at least to me in the first volume of *Population and Development Review* which was published in 1975, Nathan Keyfitz published an article entitled "How do we know the facts of demography?" Keyfitz argued that empirical relationships based on regression analysis can be misleading as they depend heavily on the cases for which data are available. If you change the sample, then estimates of regression equation will also change. Without modeling questions regarding cause and effect multiple causation and nature of relationship cannot be answered.

(Refer Slide Time: 14:00)



Illustration

Keyfitz showed that empirical relationship between percent aged 65 and over and growth rate of population was dependent on:

- (a) number of countries for which data are available; and
- (b) homogeneity among the countries.

These situations demand application of analytical and mathematical models.

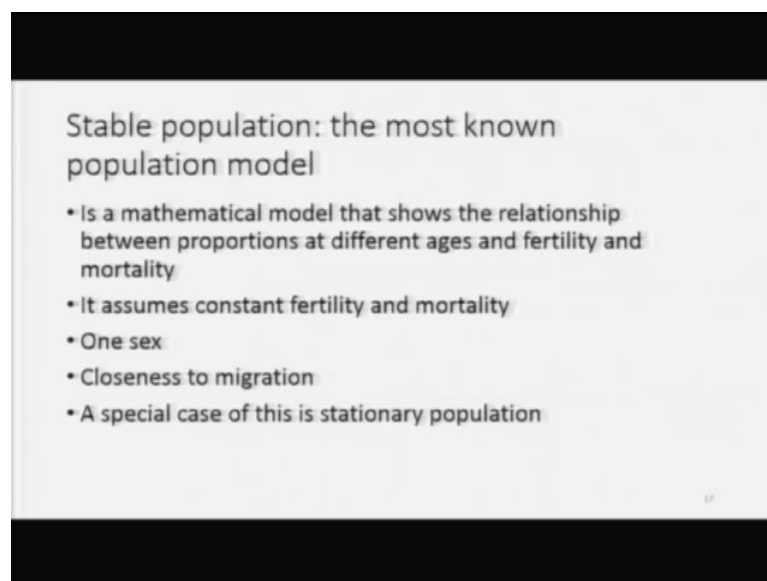
14

Illustration; Keyfitz showed that empirical relationship between percent aged 65 and over and growth rate of population, why dependent on ah: Nathan Keyfitz regressed proportion of 65 and over age dependent variable on growth rate of population as independent variable.

And he found that the regression coefficients vary and they depend on number of countries for which data are available. If you run this regression analysis on say 10 countries you find one result and if you run this integration analysis on 90 countries, then the results will be different and homogeneity among the countries. So, regression coefficient also depends on whether countries are homogeneous or heterogeneous. These situations demand application of analytical and mathematical model because the regression equation does not work. The results of regression equation will be highly variable.

One of the most known population model is a stable population model.

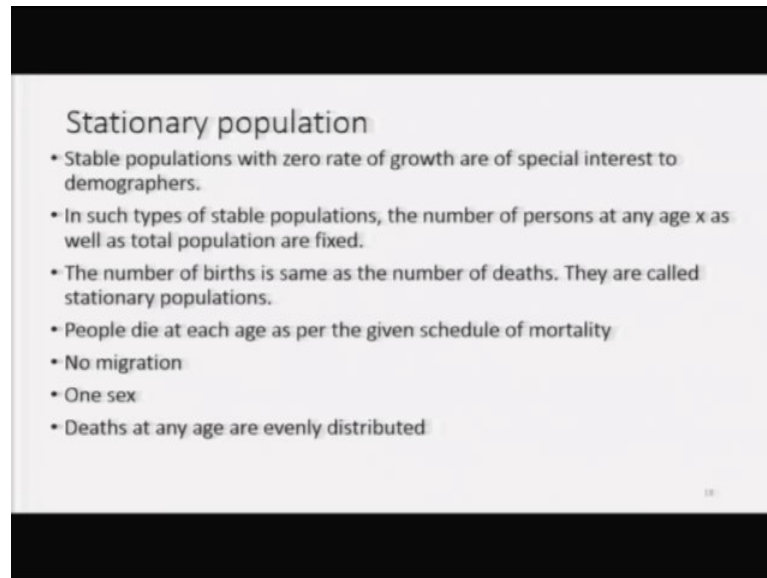
(Refer Slide Time: 15:16)



It is a mathematical model that shows the relationship between proportions at different ages and fertility and mortality rates. The stable population model assumes constant fertility and mortality, it is a one sex model. The model is close to migration means it assumes that in the population that we are modeling there is no in migration or out migration and this is a special case of life table or a stationary population. Actually

stationary population is that a stable population in which growth rate is 0. So, you can treat stationary population also as a type of stable population; growth rate is 0.

(Refer Slide Time: 16:08)



The slide is titled "Stationary population" and contains a list of seven bullet points. The text is white on a dark background. The bullet points are:

- Stable populations with zero rate of growth are of special interest to demographers.
- In such types of stable populations, the number of persons at any age  $x$  as well as total population are fixed.
- The number of births is same as the number of deaths. They are called stationary populations.
- People die at each age as per the given schedule of mortality
- No migration
- One sex
- Deaths at any age are evenly distributed

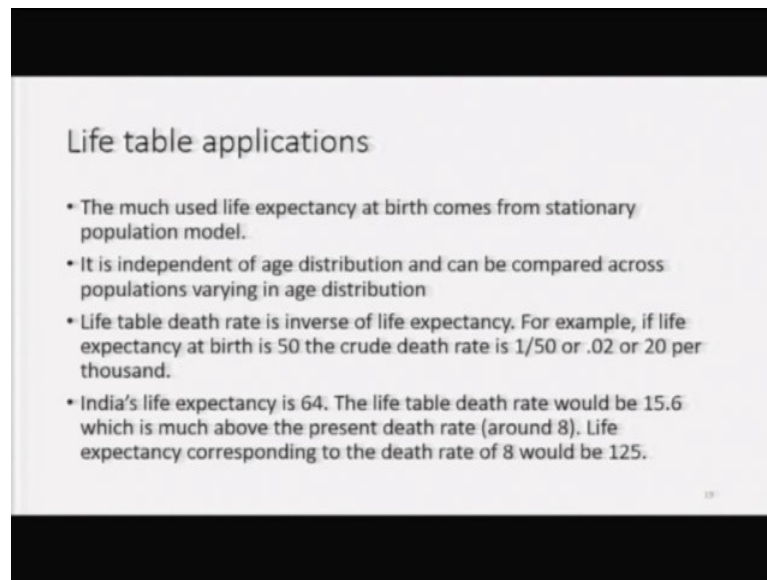
There is a small number "18" in the bottom right corner of the slide.

So, stationary population has a number of properties. A stable population with zero rate of growth are of special interest to demographers. In such types of stable populations the number of persons at any age  $x$  multiplied as well as total population are fixed. Let me repeat; in such types of stable populations means life table population or stationary population, the number of persons at any age  $x$  as well as total population are fixed. In a stationary population total population remains same and number of persons at any age also remains same. The number of births is same as the number of deaths that is why growth rate is 0. They are called stationary population.

People die at each age as for the given schedule of mortality, there is no migration. It is a one sex model, life tables or a stationary population models are built separately for males and females. They are also built sometime for country as a whole, but because of differences in age specific death rates between males and females, life tables or stationary models are used separately for males and females.

And another assumption in building stationary population model is that death at any age are evenly distributed which means that those people who will die between say age 20 and age 21; one-twelfth of them will die in the first month, one-twelfth in the second month and so on. The distribution of deaths between 20 and 21 is uniform.

(Refer Slide Time: 18:06)



### Life table applications

- The much used life expectancy at birth comes from stationary population model.
- It is independent of age distribution and can be compared across populations varying in age distribution
- Life table death rate is inverse of life expectancy. For example, if life expectancy at birth is 50 the crude death rate is  $1/50$  or .02 or 20 per thousand.
- India's life expectancy is 64. The life table death rate would be 15.6 which is much above the present death rate (around 8). Life expectancy corresponding to the death rate of 8 would be 125.

19

Applications of life tables; the much used life expectancy at birth comes from a stationary population model. It is independent of age distribution and can be compared across populations varying in age distribution. Life table death rate this is something which all of you should understand that we define life table death rate as one upon life expectancy.

For example if life expectancy at birth is 50, the crude death rate is one upon 50 or 0.02 or 20 per 1000. In India today life expectancy is around 70. So, 1 upon 70 means 0.014 or 14 per 1000 is the life table is equivalent to life table death rate in India. Now actual death rate is much lower life table death rate is 14, but actual death rate is much lower and that is because the age distribution of the actual population of India is not the same as the age distribution of life table population which is generated from age specific death rates themselves. India's life; this is this I wrote when life expectancy was smaller.

So, India's life expectancy if it is 64, then death rate would be 15.6 which is much above the death rate when life expectancy was 64 and that is due to differences in age distribution of population. And a population of a stationary model generated by lifetime by age specific death rates.

(Refer Slide Time: 19:50)

The most important of all the demographic models is the stable population model (Smith and Keyfitz, 1977; Keyfitz, 2005). It says that if individuals are born at a constant rate of 1 person per unit of time, and the survival probability of a person aged  $x$  is  $p(x)$ , then at any time the expected size of the population is given by

$$E[X] = \int_0^{\alpha} p(t) dt$$

where  $E[X]$  refers to the expected size of population at age  $x$ ,  $p(t)$  refers to chance of survival from birth to age  $t$ , and  $\alpha$  refers to the upper limit of the age distribution.

10

The most important of all the demographic models is the stable population model. It says that if individuals are born at a constant rate of one person per unit of time and the survival probability of a person aged  $X$  is  $p(x)$  then at any time the expected size of the population is given by  $E[X]$  means expected value of  $X$  equal to integral 0 to  $\alpha$   $p(t) dt$ . Where  $E[X]$  refers to the expected size of population at age  $X$ ,  $p(t)$  refers to chance of survival from birth to age  $t$  and  $\alpha$  refers to the upper limit of the age distribution which may be around 100.

(Refer Slide Time: 20:52)

Stable population model is a deterministic model and population following the stable model always results in a population in which the proportion of persons at any age (x) does not change with time. Further, suppose the individuals are born at a constant rate. Then the size of population at time t:

$$V(t) = B e^{\alpha t} \int_0^{\infty} e^{-\rho x} p(x) dx$$

Here  $\alpha$  and  $\rho$  are two constants.

The stable population model is; however, a deterministic model and population following the stable population model always results in a population in which the proportion of persons at any age x does not change with time. Further suppose the individuals are born at a constant rate, then the size of population at time t would be V t equal to B, B for births e raised to power alpha t integral 0 to alpha or 0 to infinity actually, 0 to infinity because infinity can be taken as the upper limit of the age interval or longevity e raise to power minus rho x p x dx and here alpha and rho are two constants.

(Refer Slide Time: 21:42)

This population grows or declines at rate  $\rho$ .

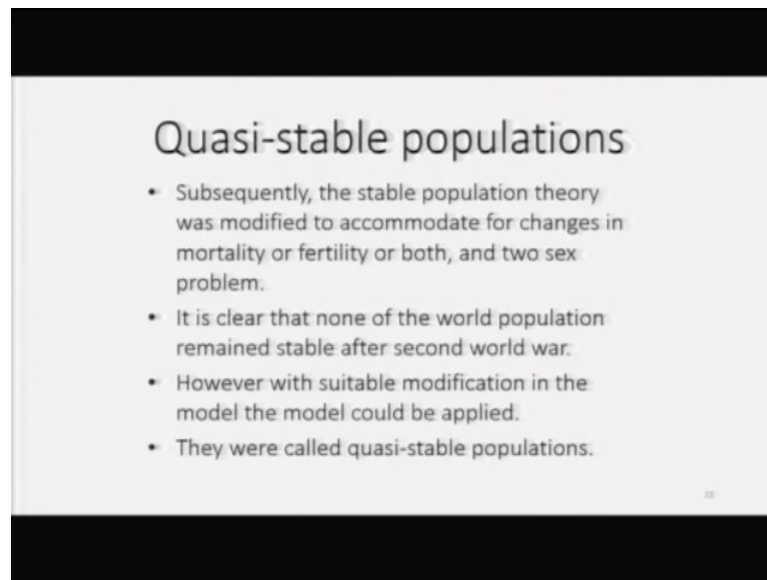
Its age distribution, i.e., proportion surviving to age x is proportional to:

$$e^{-\rho x} p(x)$$

Using the property of stable populations that age distributions of two stable populations never cross each other census growth rate and proportion of population up to a certain age (normally 35 years) were used to estimate birth and death rates for those populations which lacked reliable and complete data on them.



(Refer Slide Time: 21:51)



### Quasi-stable populations

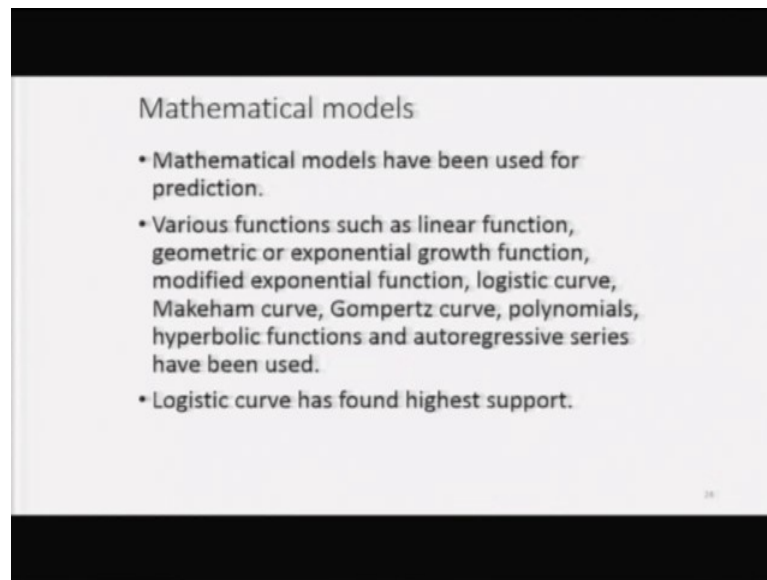
- Subsequently, the stable population theory was modified to accommodate for changes in mortality or fertility or both, and two sex problem.
- It is clear that none of the world population remained stable after second world war.
- However with suitable modification in the model the model could be applied.
- They were called quasi-stable populations.

28

Now without going into details of the stable population model, I would like to raise some other conceptual issues. Actually when we were a students, we used stable population theory to estimate birth and death rates of India, because the data that we had at that time was only from census and census gave us growth rate of population and age distribution. By choosing an appropriate stable population and this assumption that birth and death rates have been unchanging for a long time was quite safe at that time.

So, by choosing an appropriate life table for which growth rate and age distribution often signified by proportion of people below the age of 35 or below 30 or below 25, we could attribute birth rates and death rates of stable population to the real population of India. When death rates started changing means one of the two things either fertility or mortality they start changing, then we have quasi or quasi stable population.

(Refer Slide Time: 23:00)



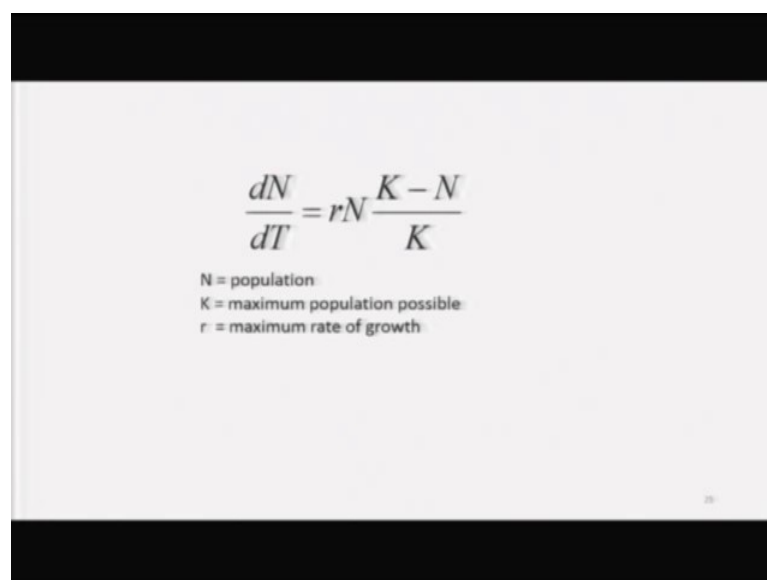
Mathematical models

- Mathematical models have been used for prediction.
- Various functions such as linear function, geometric or exponential growth function, modified exponential function, logistic curve, Makeham curve, Gompertz curve, polynomials, hyperbolic functions and autoregressive series have been used.
- Logistic curve has found highest support.

28

Mathematical models have been used for prediction and among them graphs and algebraic functions for exponential function logistic curve, Makeham curve, Gompertz curve, polynomials, hyperbolic functions and autoregressive series often used by economists are also used by demographers and logistic curve has been found to be of highest importance.

(Refer Slide Time: 23:30)


$$\frac{dN}{dT} = rN \frac{K - N}{K}$$

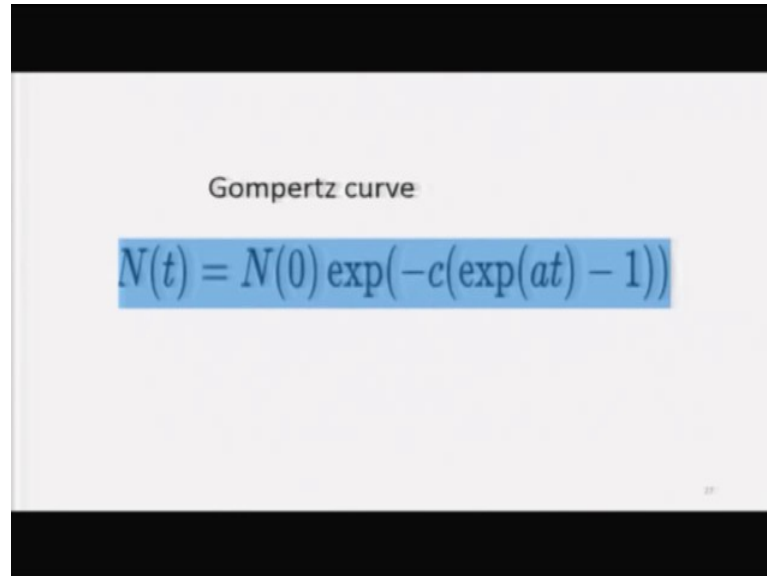
N = population  
K = maximum population possible  
r = maximum rate of growth

29

This is the equation of logistic curve  $dN$  by  $dT$  and its population  $K$  is the maximum possible population the limit that population can reach  $r$  is the maximum rate of growth.

So,  $\frac{dN}{dt}$  equal to  $r \frac{N}{K} \left( 1 - \frac{N}{K} \right)$  or  $N$  becomes size of population becomes  $K$  upon  $1 + B e^{rT}$ .

(Refer Slide Time: 23:58)

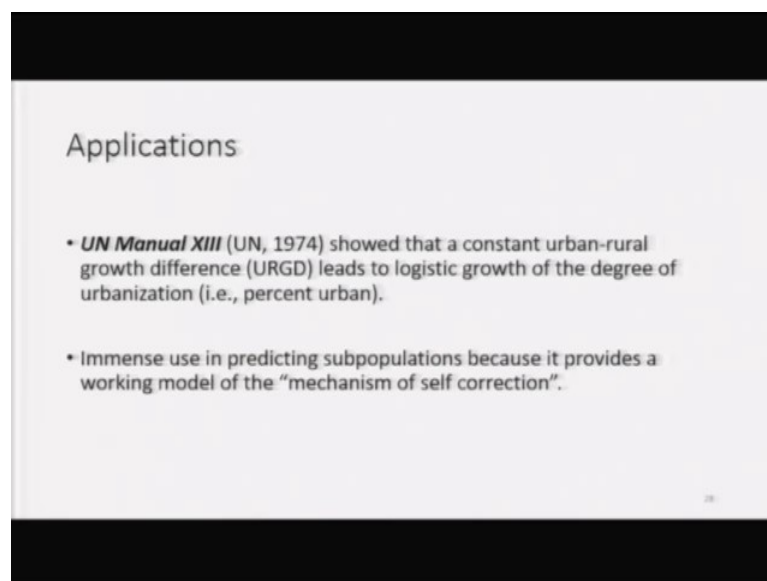


Gompertz curve

$$N(t) = N(0) \exp(-c(\exp(at) - 1))$$

This is Gompertz curve those who are more interested into curves and models, you know they can go through say, there are now several books on mathematical demography and they can also use B.D. Mishra's an Introduction to Population Study; all these curves and graphs are given.

(Refer Slide Time: 24:22)

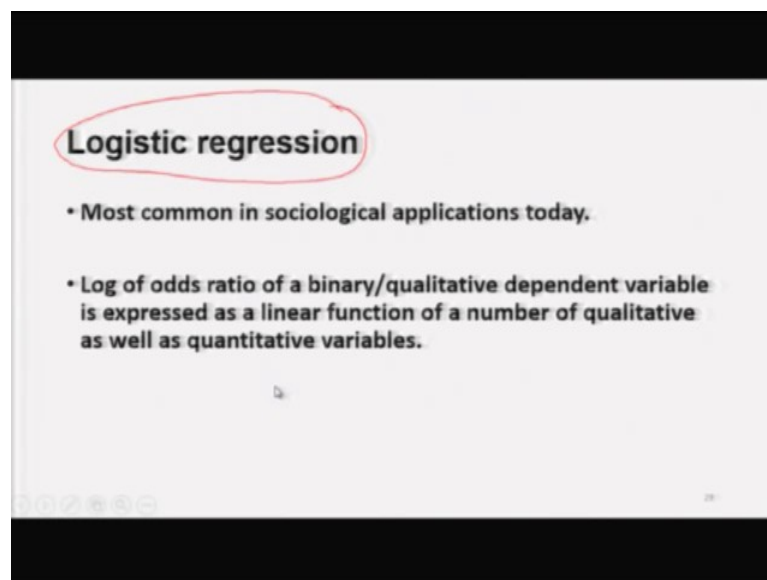
- 
- Applications
- *UN Manual XIII* (UN, 1974) showed that a constant urban-rural growth difference (URGD) leads to logistic growth of the degree of urbanization (i.e., percent urban).
  - Immense use in predicting subpopulations because it provides a working model of the "mechanism of self correction".

Among applications of these models, one can look at for simple applications of models. UN Manual XIII, which was published in 1974 and that showed that a constant urban-rural growth difference leads to logistic growth of the degree of urbanization or percent urban. Means if urban and rural growth rates separately may change, but as long as urban rural growth rate is fixed; it can be derived that then percentage urban follows a logistic model.

Initially the level of urbanization is zero as time passes percentage urban starts increases and it increases at increasing rates, a time comes when there is highest rate of growth of urbanization after that rate of growth of urbanization start declining, but percentage urban keeps on increasing.

Now this model has been of immense use in predicting subpopulations like urban population, rural population, district wise population, population of certain occupations because it provides a working model of the mechanism of self correction and it is a kind of ratio method. But more sophisticated than predicting ratio on the basis of fitting some linear or some polynomial curve to the past data; it has a logic of its own.

(Refer Slide Time: 26:58)



It is most commonly used model in sociological applications also today, then it leads to what is written here logistic regression. Logistic regression is a technique of regression analysis when the dependent variable is binary and you can code its values as 0 and 1 only. So, then for p you cannot use a simple linear regression, then we go for logistic

regression and in place of working with  $p$  we work with odds ratio is  $p$  upon one minus  $p$  and we assume that  $\log$  of odds ratios;  $\log p$  upon one minus  $p$  follows a linear regression and can be expressed in terms of socio economic or cultural variables.

So, this bullet  $\log$  of odds ratio of a binary qualitative dependent variable is expressed as a linear function of a number of qualitative or a quantitative variables. It can accommodate both qualitative and quantitative independent variables. So, the logistic model and logistic regression binary; logistic regression can they are of immense value and logistic regression interestingly can be used not only when dependent variable is divided into two categories only and you code them 0 and 1. Even when you have three or four categories and there is an order between those categories some ranking, some hierarchy then also logistic model can be used.

So, in this lecture I just wanted to introduce the purpose of modeling. There are some questions which cannot be answered by using data from experiments, from surveys, from ethnography, from case history case history method or extended case method or other method that usually sociologists use. And in that case building up mathematical models or statistical models can be of immense help. Mathematical models are usually analytical while statistical models are those which contain or which deal with probabilistic variables random variables or which include a random term called error and accordingly, then stochastic models are used.

Thank you.