MODULE TWO

BIOMETRY

DR. KOFI OWUSU-DAAKU

POPULATION GENETICS AND EVOLUTION

LECTURE IX

MEASURES OF CENTRAL TENDENCY

DR. KOFI OWUSU-DAAKU

WED MARCH 15, 2017

- The term 'measures of central tendency' refers to the determination of mean, mode and median.
- The properties of large collected statistical data are difficult to understand without further treatment.
- The vast statistical data are condensed in such a way that the basic character of the data does not change

Measures of Central Tendency Cont'd

- An average reduces the large number of data/observations to one figure.
- The average is a number indicating the central value of a group of observations.
- The average value of any characteristics is the one central value around which lie other observations.

Measures of Central Tendency Cont'd

- Thus, 'average' is a general term that describes the centre of observations.
- Three common types of average are mean, median and mode

Mean

- Mean is the sum of all observations divided by the number of observations.
- It is the most common measure of the central tendency.
- It is the best known and most useful form of average.
- The method of calculation of arithmetic mean depends upon the nature of data available, which may be explained as follows:

Mean cont'd

(a) When the observations are small in size (i.e., in a series of individual observations):

$$\text{Mean or } \overline{X} = \frac{\Sigma x}{N}$$

where X = Data (values of variable) and

N = Number of observations.

(b) Calculation of arithmetic mean in a discrete series:

$$\text{Mean or } \overline{X} = \frac{\Sigma fx}{N} \text{ or } \frac{\Sigma fx}{\Sigma f}$$

where f = Frequency,

x = The concerned variable and

N or Σf = Total number of observations.

Mean cont'd

(c) Calculations of arithmetic mean in a continuous series:

$$\text{Mean or } \bar{X} = \frac{\Sigma fm}{N}$$

where

Σfm = Total of the frequency of each class multiplied with the mid value of respective class and

N = Total of the frequencies.

Median

DR. KOFI OWUSU-DAAKU

- When all the observations of a variable are arranged in either ascending or descending order, the middle observation is known as median.
- Median is neither based on the total nor is it affected by the extreme values of variables.
- Median is a point, not a score or any particular measurement.

Calculation of Median

(i) For individual observations:

Median or M = Sum of the N + 1/2th item where N = Number of items.

(ii) For Discrete Series

For calculating median in a discrete series, frequency is made cumulative and then median is calculated on the basis of above formula.

Calculation of Median cont'd

(iii) For Continuous Series

After making the frequencies cumulative, the median item is found out as N/2th item and then the median is calculated as per the following formula:

Calculation of Median cont'd

$$M = l_1 + \frac{i}{f}(m - c)$$

or

$$M = l_1 + \frac{\frac{N}{2} - c}{f} \times i$$

where M = Median,

Ii = Lower limit of the median class,

i = Class interval of the median class,

f = Frequency of the median class,

c = Cumulative frequency of the class preceding the median class and

m = N/2th item.

MODE

- The mode may be defined as the observation with the highest frequency.
- This is a value that occurs most frequently in a statistical distribution.
- Normally, mode is frequently used for categorical data.
- Mode - Median = 2 (Median - Mean) or
- Mode = Mean = 3 (Median - Mean)

Calculation of Mode

- For individual series - After converting the data into discrete series, the modal item should be picked up as the most occurring value.
- For discrete series - Mode can be located simply by inspection of the series, i.e., the size having the highest frequency will be mode of that series.

Calculation of Mode cont'd

- For continuous series - In a distribution of grouped data, the mode is estimated at the midpoint of the class interval having the greatest frequency.
- Mode can be calculated in a continuous series by the following formula

Calculation of Mode cont'd

$$\text{MO} = i_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

or

$$\text{MO} = \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times i$$

where Ii = Lower limit of the modal class,

f_1 = Frequency of the modal class,

f_0 = Frequency of the class preceding modal class,

f_2 = Frequency of the class following the modal class and

i = Class interval of the modal class.

DISTRIBUTION

- Frequency distribution is of two types, viz., observed frequency distribution and expected frequency distribution.
- Observed frequency distribution is prepared on the basis of actual data, whereas expected frequency distribution is a theoretical one.

Distribution cont'd

- Calculations of theoretical distribution are useful in many ways such as to understand the risk and uncertainty in any event, helps in forecasting, serves as benchmarks for comparison, etc.

Types of Distribution

- There are different types of theoretical frequency distribution, but the following three are of great importance:

1.Binomial Distribution

- It is also known as Bernoulli's distribution.
- It is identified by the number of the observations, n, and the probability of occurrence which is denoted by p.

Binomial Distribution Cont'd

- The essential features of this distribution are as follows:

(a) The number of trials is fixed.

(b) There are two mutually exclusive possible outcomes of each trial.

(c) The trials are independent

Binomial Distribution Cont'd

The binomial distribution is used when a researcher is interested in the occurrence of the events and not in its magnitude.

- This distribution is widely used in industries for quality control.

2. Poisson Distribution

- Poisson distribution was developed by the French mathematician, Simeon Denis Poisson (1837).
- It is a very useful probability distribution.
- Poisson distribution gives the idea of probability of rare events, i.e., the number of trials is very small and the probability of success is also very small.

2. Poisson Distribution Cont'd

- Poisson distribution is a discrete distribution with a single parameter, i.e., the mean of distribution.
- It is widely used in insurance, spread of diseases, physiology and genetics.

3.Normal Distribution

- The pattern of distribution of data that follows the bell-shaped curve is known as normal distribution.
- Normal distribution was used by mathematicians de Moivre and Laplace in the 1700s. German mathematician and physicist, Karl Gauss, used it to analyze astronomical data, hence it is also known as Gaussian distribution.

3.Normal Distribution cont'd

- All normal distribution is symmetric.
- Normal distribution is the most useful theoretical distribution for continuous variables.
- The shape of normal distribution resembles the bell, so sometimes it is also referred to as the bell curve.
- It is the most frequently used of all probability distributions

- The general equation that describes normal distribution curve is as follows:

$$Y = \frac{N}{6\sqrt{2\pi}} - e\frac{x^2}{26^2}$$

where N = Number of measures,

Y = Frequency,

π = 3.1416,

c = 2.7183,

σ = Standard deviation of the distribution and

x = Deviation of any unit of measurement from the mean.

- Biological distribution is generally assumed as normally distributed.

CORRELATION

- Correlation is a statistical technique showing relationship between two variables.
- It is one of the most common and most useful statistics.
- The possible correlations range from +1 to - I.
- A zero correlation indicates that there is no relationship between variables.

Correlation cont'd

- A correlation of - 1 indicates that if one variable increases the other decreases, while a correlation of + I indicates that both variables move in the same direction.
- Further, it shows the closeness or degree of relationship between the variables.
- Correlation is also a marker of interdependence between two variables

Types of Correlation

- On the basis of nature of relationship between the variables, correlation may be of the following types:

(a) Positive or negative

(b) Simple, partial or multiple

(c) Linear or non-linear

Degree of Correlation

- On the basis of coefficient of correlation, the degree of correlation may be of the following types:

(a) Perfect ,(b) Limited and (c) Absent

- The degree of relationship between two variables is the coefficient of correlation represented by the symbol 'r'.
- It is called Karl Pearson's coefficient of correlation and is most widely used.

Calculation of Karl Pearson's Coefficient of Correlation

- For individual series

$$rxy = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$$

where

r = Correlation coefficient,

x = Deviation of X variables $(X - \bar{X})$ and

y = Deviation of Y variables $(Y - \bar{Y})$.

y = Deviation of Y variables (Y − Ȳ).

Product-Moment Correlation

- When number of observation (N) is small, their correlation can be calculated by product-moment method, according to the formula given below:

$$\frac{\Sigma xy - \frac{\Sigma x \Sigma y}{N}}{\sqrt{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{N}\right] \times \left[\Sigma y^2 - \frac{(\Sigma y)^2}{N}\right]}}$$

$$\frac{\Sigma fdxdy - N\left(\frac{\Sigma fdx}{N}\right)\left(\frac{\Sigma fdy}{N}\right)}{\sqrt{\left[\frac{\Sigma fdx^2}{N} - \left(\frac{\Sigma fdx}{N}\right)^2\right] \times \left[\frac{\Sigma fdy^2}{N} - \left(\frac{\Sigma fdy}{N}\right)^2\right]}}$$

REGRESSION ANALYSIS

- Regression analysis is the technique for the prediction of the relationship of a particular variable with another, on the basis of its relationship with a third variable.
- The variable to be estimated is called the dependent variable and the variable that provides the basis for estimation is called the independent variable.

Regression Analysis cont'd

- In multiple regression, there are two or more independent variables and one dependent variable.
- In other words, from regression analysis, we can estimate the value of one variable from the given value of the other variable. For example, we can find out the expected weight of a fish from a given length.

Regression Analysis cont'd

- The relationship between the independent variable (X) and the dependent variable (Y) is expressed regression equation. The regression equation expresses the regression lines.

Regression Analysis cont'd

- Since there are two regression lines, there are two regression equations.
- The regression equation X on Y shows the variation in the values of X for changes in Y.
- Likewise, regression equation Y on X describes the variation in the values of Y for changes in X.

Regression Analysis cont'd

- Regression equation:
- X = *a* + *by (x* on *y)*
  Y = *a* + *bx* (y on x)
- where 'a' is a constant (the point where regression line touches (Y-axis) and 'b' is also a constant call regression coefficient.

Regression Analysis cont'd

- The multiple regression equation shows the effect of a number of independent variables at the same time which may be written as follows:
- $Yc = a + b_1x_1 + b_2x_2 + b_3x_3$.
- Where Y c = Value of dependent variable

  $x_1, x_2, x_3 \ldots\ldots$ = Independent variable

  $b_1, b_2, b_3 \ldots$ = Regression coefficient

HYPOTHESIS TESTING AND TEST OF SIGNIFICANCE

- The test of significance is used by the researchers to determine whether the difference between calculated value and the hypothetical parameter is significant or not.
- It establishes whether there is relationship between variables or the observed values have been produced by the chance.

Hypothesis Testing And Test Of Significance Cont'd

- The phrase test of significance was coined by R A Fischer (1925).
- Every test of significance is associated with a basic concept known as the hypothesis.
- The hypothesis is basically a statement about the population parameters.
- It can be grouped into two types, viz., null hypothesis and alternative hypothesis

Hypothesis Testing And Test Of Significance Cont'd

DR. KOFI OWUSU-DAAKU

REVIEWED
MARCH 15, 2017

- Statistical inferences are drawn on the basis of information we get from the sample
- In other words, it is possible to make reasonable estimates from the sample data available.
- Even if we don't know about a population, we can get reliable information about it on the basis of random sample from that population

Hypothesis Testing And Test Of Significance Cont'd

- The estimation deals with the methods by which population parameter/characteristics are estimated from sample information, whereas hypothesis testing deals with the process involved in the acceptance or non-acceptance of the assumption or a statement about the population parameter.

Hypothesis Testing And Test Of Significance Cont'd

- Hypothesis testing enables us to verify whether or not such statements are in agreement with the available data.

Null and Alternative Hypothesis

- The hypothesis to be tested is called 'Null Hypothesis' and is represented by Ho. This may be written as follows:
- Ho: μ -x= 0 [where x = Sample mean and μ = Population mean]
- From the above equation, it can be concluded that there is no difference between the population mean and sample mean.

Null and Alternative Hypothesis cont'd

- Null hypothesis must be tested.
- To test the null hypothesis, there is an alternative hypothesis represented as HI.
- If this alternative hypothesis is correct, the null hypothesis is rejected.

Errors in Testing of Hypothesis

- Since the acceptance or rejection of null hypothesis (Ho) depends on sample study, there is every chance of error. The error may be:

I. Type I *(α)* error - to reject null hypothesis when it is true.

2. Type II (β) error - to accept null hypothesis when it is false.

Level of Significance

- The probability of committing *a* -error is called level of significance.
- 5 per cent (0.05) and 1 per cent (0.01) are the most commonly used levels of significance.
- 5 per cent level of significance shows that out of 100 times, there is a probability that 5 times correct Ho will be rejected.

Test of Significance

- An assessment of significance of difference between parameters of different samples is known as the test of significance.
- Such a test gives an idea whether observed differences between two samples are significant or have occurred due to chance.

STANDARD ERROR OF MEAN

- The standard deviation of the sample means is called the standard error of mean.

- The standard deviation of the sample means is called the standard error of mean.

$$\text{S.E. } \bar{X} = \sqrt{\frac{\Sigma(\bar{x} - \mu)^2}{N - 1}}$$

where μ = Mean of the sample mean and X = Sample mean.

- When the standard deviation of a population is known,

$$\text{S.E. of mean} = \frac{SD}{\sqrt{N}}$$

Standard Error Of Mean (Sex) Cont'd

- A small value of standard error of mean is a clear indication of the fact that the various values of X are close to each other and average difference between these Xs and μ is small.
- As the sample size increases, the standard error of mean becomes smaller. At the same time, on increasing the sample size, various sample means become more uniform.

Standard Error Of Mean (Sex) Cont'd

- Standard error is useful in testing a given hypothesis.
- It gives an idea about unreliability of a sample.

STANDARD ERROR OF STANDARD DEVIATION

- Standard deviation of different samples of the same population varies.
- So the standard error of standard deviation can be calculated to test the significance.
- SE of standard deviation data can be calculated as follows:

STANDARD ERROR OF STANDARD DEVIATION

DR. KOFI OWUSU-DAAKU

- SE of standard deviation data can be calculated as follows:

$$\text{SE}\,\sigma = \frac{SD}{\sqrt{2N}}$$

For grouped data, it can be calculated by the following formula:

$$\text{SE}\,\sigma = \frac{SD}{\sqrt{2\Sigma f}}$$

STUDENT t-TEST

- Student t-test is a small sample test.
- Student t-test was developed by W S Gosset (1908).
- Gosset published his work in pseudonym 'Student' in 1908.
- It is the most common statistical technique used to test the hypothesis based on difference between sample means.

It is also called t-ratio because it is a ratio of difference between two means.

$$t = \frac{\overline{X} - \mu}{S / \sqrt{N}}$$

where X = mean of the sample,

S = Standard deviation of the sample and

N = Number of observations in the sample.

STUDENT t-TEST Cont'd

- A conclusion based on t-test is good if the distribution is normal or near and samples are chosen randomly.
- Fisher's table gives the highest obtainable values of 't' under different probabilities, with (P) in decimal fractions corresponding to the degrees of freedom.

STUDENT t-TEST Cont'd

- Probability of occurrence of any calculated value of 't' is determined by comparing it with the value given in the table.
- If the calculated 't' value exceeds the value given in the table, it is said to be significant.

Application of the t - Test

(a) Student 't' test for single mean is used to test a hypothesis on specific value of the population mean.
(b) Student t-test is used to test the difference between the means of two samples.

(c) The paired t-test is applied when the two samples are dependent.

(d) A t-test is used to test the significance of an observed correlation coefficient.
(e) A t-test is used for testing significance of regression coefficient.

CHI-SQUARE (x^2) TEST

- Chi-square test is the most commonly used method for comparing frequencies.
- **It** is a statistical test that is used to measure difference between an observed data with the data we would
- expect according to a given hypothesis.
- Chi-square is calculated on the basis of frequencies in a sample.

CHI-SQUARE (x^2) TEST CONT'D

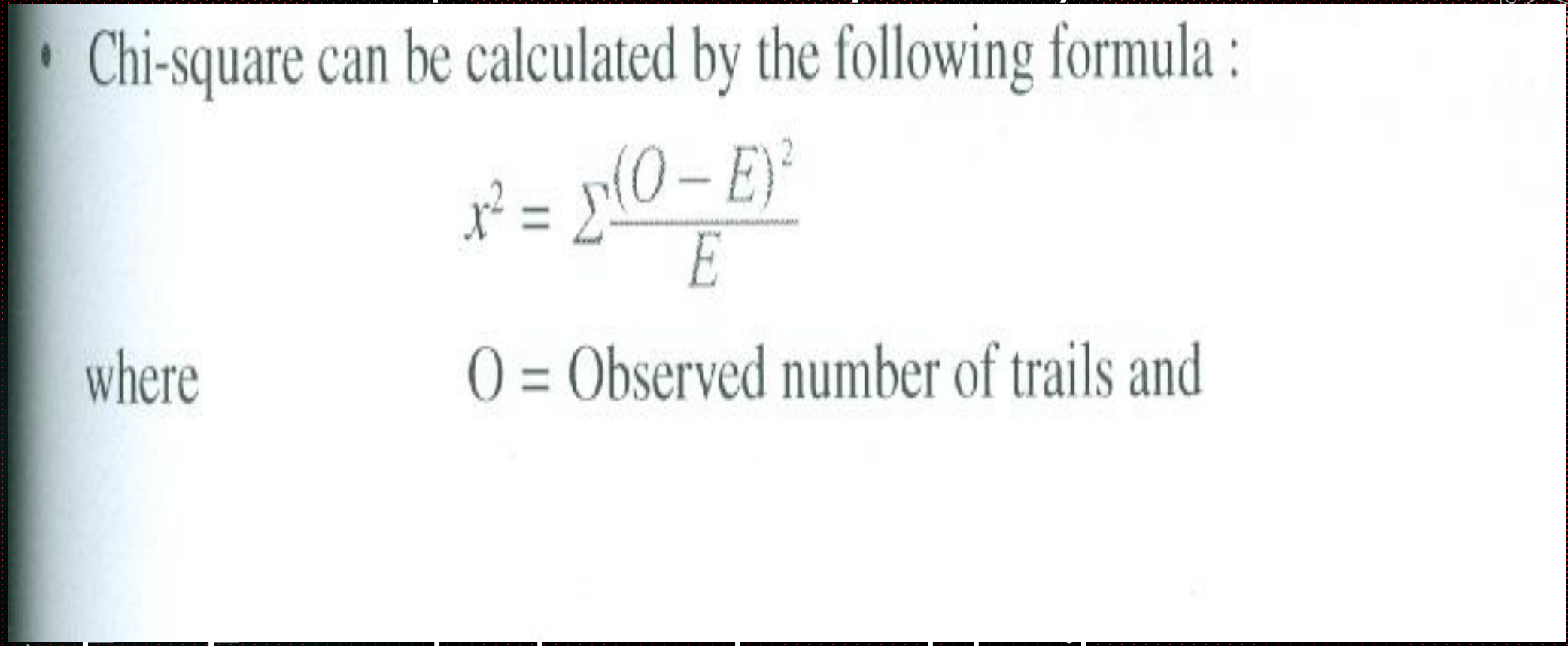
- **It** is used as a test of significance when the data are in forms of frequencies or percentages or proportions.
- **It** is one of the simplest and widely used non-parametric tests in statistical analysis.
- Chi-square test compares the observed value with the expected value and find out how far the differences between the two values can be attributed to fluctuations of simple sampling.

CHI-SQUARE (x^2) TEST CONT'D

- The Chi-square test was developed by Prof. A R Fischer (1870) and it was further developed by Karl .
- Pearson (1906) in its present form.
- The following are the essentials to apply x^2 test:
- Random sample
- Qualitative data

CHI-SQUARE (x^2) TEST CONT'D

- Lowest expected frequency not less than 5



- E = Expected number of trails.

CHI-SQUARE (x^2) TEST CONT'D

- In x^2 test, the number of degrees of freedom is equal to the number of classes minus one.
- The value of x^2 depends on the degrees of freedom.
- x^2 test is also applied as a test of goodness of fit as it shows the closeness of observed and expected frequency.

Characteristics of x^2 test

(a) It is based on frequencies.
(b) It is non-negative.
(c) It is highly skewed.
(d) It is based on degrees of freedom.

Characteristics of x^2 test

(e) With the change in degree of freedom, a new chi-square distribution is created.

(f) The shape of chi-square distribution does not depend on the size of sample. It may depend upon the number of categories.

Uses

(a) A chi-square test is used as a test of homogeneity. It is a test which is used to determine whether several populations are similar or equal or homogenous in some characteristics.

(b) Chi-square test is used as test of independence. With the help of chi-square test, one can be able to know whether two attributes are associated or not.

Uses Cont'd

(c) Chi-square test as a test of goodness of fit is used to determine whether the sample data are in consistent with the hypothesized data.

ANALYSIS OF VARIANCE

- To test the hypothesis whether the means of several samples have significant differences or not, a method called analysis of variance is used.
- This method is based on the comparison of variances estimated from various sources.
- The analysis of variance is based on the following assumptions:
(a) Populations are normally distributed.

ANALYSIS OF VARIANCE CONT'D

- (b) Populations from which the samples have been taken have means (μ_1 , μ_2 , μ_3 etc.) and variances ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ .....)
- (c) Samples have been randomly selected.
- F-test is widely used in the analysis of variance and is calculated as follows:

ANALYSIS OF VARIANCE CONT'D

- F = *variance between samples / variance within samples*

The analysis of variance is mainly of the following two types:

(i) One-way analysis of variance

(ii) Two-way analysis of variance

I. One-way Analysis of Variance

- Here, analysis of variance observations are grouped on the basis of single criterion, i.e., the influence of only one factor is considered.
- In this type of analysis of variance, samples have been taken from normal populations with common variance.

II. Two-way Analysis of Variance

DR. KOFI OWUSU-DAAKU

- Here we have to take consideration of the influence of two factors.
- The data are grouped according to the two different factors