**BIOSTATISTICS LEVEL II NOTES 2014** 

****COMPILED BY EFFIE NAILA

**OUTLINE**

INTRODUCTION TO BIOSTATISTICS 



PRESENTATION OF DATA 



CALCULATING MEASURES OF DISPERSION IN 

GROUPED DATA

**1. INTRODUCTION TO BIOSTATISTICS **BY: ERASTUS NJERU

**OBJECTIVES** A. Definitions 

B. Types of data 

C. Descriptive statistics for quantitative & qualitative data 

•Measures of central tendency & dispersion •Data presentation (frequency tables & graphs)

**A. DEFINITIONS** • Statistics: 

• This is the science concerning collection, organization & summarization of data as well as interpretation and drawing inference.

• Statistics are the summary of the indices for data obtained from a sample. • Biostatistics:

• Application of statistical methods in biological sciences.

• It is divided into 2 parts:

1. Descriptive: organization & summarizing

2. Inferential: drawing inference

**IMPORTANCE OF BIOSTATISTICS** • The main interest of public health professionals is to improve the health status of the population through informed decision making. 

• The public health professional should be able to translate data into meaningful information usable as evidence for public health decisions, ideally policies.

**BIOSTATISTICS EQUIPS WITH SKILLS FOR:** 

• The management and analysis of health data

• Critical appraisal of available health literature and new findings

• Sharing new information with colleagues and policy makers • Preparation of information to lobby for material support and translating knowledge to policy

**KEY CONCEPTS** • Population: collection of **all** items/ subjects **of interest** 

• Sample: part of the population selected to represent the population. Could be: 

• Random 

• Non – random 

• Datum: raw fact/ information collected on one individual of interest, on a variable of interest. 

• Data: facts on two or more individuals

**CONT.** • Parameters: summary of indices for describing the entire population 

• Statistics: summary indices for data obtained from a sample • Variables: characteristics that vary from subject to subject or from time to time. 

• Some variables change frequently while others don’t change as often

**TYPES OF VARIABLES** • Can be classified in accordance to the following contexts: • In data collection/ presentation (descriptive statistics) • How the values are measured/ observed (scale of measurement) 

• In data analysis (inferential statistics)

• Whether they are a response or explanatory variable, i.e., the purpose of the variables.

**THE ABOVE CAN ALSO BE** 

**CHARACTERIZED AS FOLLOWS**

• Discrete variables 

• Can take only certain values and none in between • E.g. the number of patients in a hospital census • Continuous variables 

• May take any value (typically between certain limits) • Most biomedical variables are continuous. • E.g. patient’s weight, height, age, BP etc.

**B. TYPES OF DATA: Q/Q CLASSIFICATION**

QUANTITATIVE: 

• The values of a quantitative variable can either be: 

• **Discrete**: characterized by gaps (interruptions) in the values that it can assume e.g. number of lesions, number of children etc.

• **Continuous**: can assume any value on the real number line e.g. various measurements on individuals such as weight, age, gestation age etc.

QUALITATIVE: 

• The values of a qualitative variable divide the population into categories e.g. gender

**DATA WILL ALWAYS FORM ONE OF FOUR SCALES OF MEASUREMENT** 

**(STEVEN’S CLASSIFICATION)**

**NOMINAL/** 

**CATEGORICAL** e.g. gender, race, religion

• The data are in form of labels dividing the population into 

qualitative categories • There is no order between the various values of the variables • Nominal data that fall into only 2 groups are dichotomous data.

**ORDINAL** e.g. severity of edema, position in a race 

• The data are in form of ranks 

• There is meaningful order between the values of the variables

• There is no sensible arithmetic difference

**INTERVAL** e.g. 

temperature, calendar dates, social class & personality scales

• The data are usually in form of measurements on some scale 

• They have order between the values of the variables

• There is a sensible arithmetic difference • The zero is arbitrary and therefore ratios of scores are not

meaningful.

**RATIO** e.g. age, weight, height 

• Same properties as interval scale data 

• The zero is absolute • Most biomedical

variables lie here

• The kelvin scale is the only ratio scale of

temperature

**2. PRESENTATION OF DATA** 

A.PICTORIAL PRESENTATION 

B.NUMERICAL PRESENTATION

**KEY CONCEPTS** • Methods of pictorial presentation of data: 

• Tables: 

• Frequency distribution table 

• Contingency table 

• Graphs: 

• Line graph, Bar chart, Pie chart, Histogram, Frequency polygon, Ogive, Stem - & - leaf, Box plot & Scatter diagram • Measures of location & variability

**A. PICTORIAL PRESENTATION OF DATA: FREQUENCY DISTRIBUTION TABLE** 

• In a frequency distribution table there is a list of all the possible values in descending order with a record of the frequency (ƒ) of each. 

• Grouped frequency distributions: 

• The individual scores are grouped with each group encompassing an equal class interval. 

• Relative frequency distribution 

• Shows the percentage of all the elements that fall within each class interval. • Relative frequency = ƒ in that class interval 

nX 100

**CONT.**

| **AGE (MONTHS)** 15 – 19  | **FREQUENCY** 5  | **RELATIVE FREQUENCY** 20% |
| --- | --- | --- |
| 20 – 24  | 6  | 24% |
| 25 – 29  | 8  | 32% |
| 30 – 34  | 4  | 16% |
| 35 – 39  | 2  | 8% |
| TOTAL  | 25  | 100% |

Age distribution of children seen, center X , August 2004

**OGIVE (CUMULATIVE FREQUENCY POLYGON)** 

• This shows the percentage of elements lying 

within and below each class interval.

• Typically forms a characteristic S – shaped 

curve, i.e., ogive.

**GRAPHICAL PRESENTATIONS OF FREQUENCY DISTRIBUTIONS** 

• Frequency distributions are often presented as graphs, most commonly as histograms. 

• In the **histogram** the abscissa (X or horizontal axis) shows the grouped scores and the ordinate (Y or vertical axis) shows the frequencies. 

• To display nominal data, a **bar graph** is typically used. 

• A bar graph is identical to frequency histograms except that each rectangle on the graph is separated from the others by a space, showing the data form discrete categories.

**CONT.** • For ratio or interval scale data, a frequency distribution may be drawn as a frequency polygon in which the mid points of each class interval are joined by straight lines.

**BAR & PIE CHART**

**HISTOGRAM** 

**WHAT IS IMPORTANT:**

• Skewness: 

• Measure of the asymmetry (departure from symmetry) of the distribution • Left skewed data are characterized by a pile – up of observations to the right (long left tall) 

• Also referred to as a negative skew

• Right skewed data are characterized by a pileup of observations to the left (long right tail) 

• Also referred to as a positive skew 

• Kurtosis – degree of peakedness

**CONTINGENCY TABLE** 

| SINGLE  | **MALE** 9 (45%)  | **FEMALE** 20 (40%)  | **TOTAL** 29 (41.4%) |
| --- | --- | --- | --- |
| MARRIED  | 6 (30%)  | 25 (50%)  | 31 |
| DIVORCE SEPARATED  | 3 (15%) 2 (10%)  | 4 (8%) 1 (2%)  | 7 3 |
| TOTAL  | 20 (100%)  | 50 (100%)  | 70 (100%) |

Distribution of subjects attending clinic X by marital status & sex, Dec 2013

**LINE GRAPH**

**CONT.**

**FREQUENCY POLYGON**

**STEM - & - LEAF PLOTS** • Use digits in observation to give picture 

• 1st set of digits 🡪 stem 

• 2nd set 🡪 leaves 

• Age of children example:

• Data have only 2 digits:

Stem 1 

2

3

Leaf 97 

8625 20

**PERCENTILES** • The pth percentile is a value such that p% of the observations fall below that value. 

• Example 1: 10% of data are less than the 10th percentile • Example 2: 50th percentile = ? 

• Q1 = lower quartile (25% percentile)

• Q3 = Upper quartile (75% percentile)

**B. NUMERICAL PRESENTATION OF DATA** • 2 aspects of data are important in presentation/ description: • Location 

• On the scale of observations, where do our observations lie? • Also known as **central tendency or average** 

• Variability

• How far are the observations from one another? 

• Also known as **spread or dispersion** 

****• This is important ins statistical inference.

**I. MEASURES OF LOCATION/ CENTRAL TENDENCY: ARITHMETIC MEAN** 

• Sum of all observations divided by the number of observations. 

• It is symbolized by a {�� } in the sample and a {*µ*} in the population. 

• It is most amenable to mathematical manipulations and is not suitable when there are outliers, i.e., in very skewed distributions 

• It is the measure of central tendency that best resists the influence of fluctuation between different samples.

**FORMULA FOR** Sample mean, �� = Σ*x*i �� 

Population mean, *µ* = Σ*x*i ��

**MEDIAN** • This is the figure that divides the frequency distribution in half when all scores are listed in order. 

• If the number of observation is even, calculate the arithmetic mean of the 2 middle observations to arrive at the median. 

• The median is insensitive to small numbers of extreme scores (outliers) and therefore, it is a very useful measure of central tendency for highly skewed distributions. 

• It is the same as the 50th percentile.

**MODE** • This is the observed value that occurs with the greatest frequency. 

• On a frequency polygon, it is the highest point on the curve. • If 2 scores both occur with the greatest frequency, the distribution is bimodal; if more than 2 scores occur with the greatest frequency, the distribution is multimodal. • It is not unique

**II. MEASURES OF VARIABILITY/ DISPERSION/ SPREAD** 

•Variability is the extent to which the observations are clustered together or scattered about. •There are 3 important measures of variability: •Range 

•Variance

•Standard deviation

**RANGE** •This is the difference between the largest and smallest value. 

•Not suitable when there are outliers •Mid – range = Q3 + Q1 

2

•Inter – quartile range (IQR) = Q3 – Q1

**VARIANCE & STANDARD DEVIATION** •Sample variance is defined as the sums of squares of the differences between each observation in the sample and the sample mean divided by 1 less than the number of observations.

**CONT.** • Calculation of both the variance & the standard deviation, involves the use of **deviation scores**, which are found by subtracting the 



distribution’s mean from each value. • Deviation score, *x* = X - x 



• Where {X} is a value & {x } is the mean 

• NB: the sum of all deviation scores {Σx} should be zero.

• Variance of a distribution is simply the mean of the squares of all the deviation scores in the distribution.

**FORMULA FOR VARIANCE **1. Find deviation score (x) for each

value 

2. Square each deviation score to eliminate negative signs 

3. Obtain their mean 

• NB: (*n – 1*) is the denominator for sample variance, instead of *n*. This is 

Population variance, σ2 =Σ (X − µ)2 

*N*or Σ (x)2

*N*

Sample variance, *S2*

done as the former gives a less biased 

estimate of variance of the population. 

= Σ (X − x)2 

*n − 1or* Σ (x)2 *n − 1*

****STANDARD DEVIATION** • That variance is expressed in square units of measurement, limits its usefulness as a descriptive term and hence the standard deviation remedies this problem. 

• It is the **square root of variance** and so it is expressed in the same units of measurement as the original data. 

• The standard deviation of a population is expressed as, {***σ***} and that of a sample is expressed as {***S***}

**CONT.** • The SD is particularly useful in normal distributions because the proportion of values/ elements in the normal distribution (i.e., the proportion of the area under the curve) is a constant for a given number of SDs above or below the mean. 

• Approximately 68.27% of the distribution falls within ± 1 SD from the mean. • **Approximately 95% of the distribution falls within ± 0.96 SD from the mean.** 

• Approximately 95.45% of the distribution falls within ± 2 SD from the mean. • Approximately 99% of the distribution falls within ± 2.58 SD from the mean. • Approximately 99.73% of the distribution falls within ± 3 SD from the mean.

**COEFFICIENT OF VARIATION** •Standard deviation divided by mean •S~~x~~ 

•It has no units and can be used to compare variability in 2 groups/ populations where different units are used.

**Z SCORES** • The **z score/ standard normal deviate** of any statistical observation in a normal distribution is the number of standard deviations the observation lies above or below the mean of the distribution, 

• If the observation lies above the mean it will have a positive z score and if it lies below the mean it will have a negative z score Standard deviation = Xi − µ 

• Z score, Zi = Value − Mean σ 

• i.e. for every observation Xisubtract population mean and divide by the standard deviation.

**USE OF Z SCORES** • Since they are standardized or normalized they allow scores on different normal distributions to be compared e.g. a person’s height and his weight 

• They can be used to find the proportion of a distribution that corresponds to a particular score. 

• They can be used to find the score that divides the distribution into specified proportions 

• They allow the specification of the probability that a randomly picked observation will lie below or above a particular score.

**STANDARDIZING THE NORMAL DISTRIBUTION**

****

**STANDARD NORMAL DISTRIBUTION**

**FINDING PROBABILITIES FOR Z SCORES**

**EXAMPLE** • The population of neonates in a certain hospital is known to have birth weight that are normally distributed with mean 3.2 kg and variance 2.5 kg 

• What is the probability of getting a neonate with a birth weight of > 3.6kg? 

• What is the probability of getting a neonate with a birth weight between 2.0 and 2.5 kg?

**BERNOULLI TRIALS** • This is an experiment (usually observation of an individual) which has two possible outcomes, i.e.,: 

• Y/N; 0/1; +ve or –ve; present/ absent; success/ failure • Examples: 

• Is the baby a boy?

• Does the patient have cholera? 

• Is the client positive? 

• Does the toss of coin result in ‘head’?

**BINOMIAL DISTRIBUTION** • Note that it is possible to have a Bernoulli process from a non – binary variable. 

• Ex. Does patient have severe edema? Does roll of die result in a ‘2’ ? 

• A series of Bernoulli trials (which can be viewed as a series of observations) gives rises to a binomial distribution.

**3. CALCULATION OF** 

**STATISTICS FROM GROUPED DATA** 

****BY: ERASTUS NJERU

**THE CODED DATA METHOD** 1. Calculate the cumulative frequency & estimate the group containing the median. Assign it the value 0. 

2. To get the coded data, xc = 

Mid point of group − Mid point of median group

Class Width

3. Get the coded mean, x c = Σƒxc 

Σƒ

4. Get the coded variance, Sc2 = Σƒ{xc2} − n{x c}2 

N − 1

5. Reverse the coding

**REVERSING THE CODING** • To get the mean, variance & standard deviation of the data set, consider the following: 

• If a constant ***k***, is added to all observations: 

• Mean increases by ***k*** 

• Variance doesn’t change 

• If all observations are multiplied by a constant ***k***: 

• Mean & Standard deviation are each multiplied by ***k*** 

******• Variance is multiplied by ***k***2

**THEREFORE:** •Actual mean = {Coded mean, x c X Class Width} + Mid point of median group 

•Actual variance = Coded variance X {Class Width}2 

•Actual standard deviation = Coded SD X Class Width

| **Class limits** | **Mid –** **point, (xi)** | **Freq. (ƒ) c. ƒ**  |  | **Coded** **data (xc)** | **ƒxc** | **ƒ{xc}²** |
| --- | --- | --- | --- | --- | --- | --- |
| 15 - 20  | 17.5  | 55  | 55  | -2  | -110  | 220 |
| 20 - 25  | 22.5  | 69  | 124  | -1  | -69  | 69 |
| 25 - 30  | 27.5  | 84  | 208  | 0  | 0  | 0 |
| 30 - 35  | 32.5  | 47  | 255  | 1  | 47  | 47 |
| 35 - 40  | 37.5  | 26  | 281  | 2  | 52  | 104 |
|  |  |  |  |  | -80  | 440 |
| Σƒxc ⮚ Coded mean, x c= Σƒ −80281 ⮚ ⮚ **- 0.2847** | Σƒ{xc2} − n{x c}2N − 1 ⮚ Coded variance, Sc2 =  ✔ Σƒ{xc}² 🡺 440 ✔ n(x c)2 = 281 X {-0.2847}2 🡺 22.78 417.22280 🡺 **1.4901**⮚  |
| ⮚ Actual Mean = {(- 0.2847) X 5} + 27.5 ⮚ 🡺 **26.0765** | ⮚ Actual variance = 1.4901 X 52 ⮚ 🡺 **37.2525** ⮚ Actual standard deviation = √(1.4901) X 5 ⮚ 🡺 **6.1035** |

**CALCULATING THE MEDIAN FOR GROUPED DATA** 

****1. Find the nth value of median = c. ƒ2 

2. Median =

• Lower Median Class Boundary (l.m.c.b) + 

{nth value of median − c. ƒ of previous class group 

Frequency of median class group} X Class width

| **Weight in kg for a sample of school children** Class limits Mid - point, (xi) Freq., (ƒ) ƒx {ƒx}² c. ƒ |
| --- |
| 15 – 20  | 17.5  | 5  | 87.5  | 1531.25  | 5 |
| 20 – 25  | 22.5  | 6  | 135  | 3037.5  | 11 |
| 25 – 30  | 27.5  | 8  | 220  | 6050  | 19 |
| 30 – 35  | 32.5  | 4  | 130  | 4225  | 23 |
| 35 – 40  | 37.5  | 2  | 75  | 2812.5  | 25 |
|  |  | 25  | 647.5  | 17656.25 |  |
| For calculating **mean** the assumption is that all observations in each interval lie at the mid – point of the interval: ⮚ Mean =Σƒx ~~Σƒ~~ ⮚647.5 ~~25~~ 🡺 **25.90** | Σƒ{x2} − n{x }2 ~~N − 1~~ ⮚ Variance = 17656.25 − 25 {25.90}2 ~~24~~ ⮚ 886 ~~24~~ 🡺 **36.92**⮚  |
| For calculating the median the assumption is that all observations in the interval containing the median are equidistant: ⮚25~~2~~ = 12.5th value (**corresponds to the class group whose limits are 25 – 30**) ⮚ 24.5 + {12.5 − 11 ~~8~~ X 5} = 24.5 + 0.9375 = **25.4375** |

**“IT IS NO MEASURE OF HEALTH TO BE WELL ADJUSTED TO A PROFOUNDLY SICK SOCIETY”** 

JESUS CHRIST IS THE ONLY WAY, THE ONLY TRUTH & THE ONLY LIFE.