Arthur Jarvis University, Akpabuyo.

STA 203 : STATISTICS FOR AGRIC/BIOLOGICAL SCIENCES.

Course Outline.

- *Use of Statistics in Agriculture and Biology*
- *Sampling*
- *Frequency Distribution*
- *Estimation and Hypothesis Testing*
- *Contingency Tables*
- *Correlation and Regression & Covariance.*

SECTION ONE

USE OF STATISTICS IN BIOLOGY & AGRICULTURE

INTRODUCTION:

Statistics is a familiar and accepted part of modern world that is concern with obtaining an insight into the real world by means of the analysis of numerical relationships. It is used in almost all fields of human endeavor. It is applied in sports, public health, education, surveys, operations research, quality control, estimation and prediction. This unit discusses the meaning of Statistics and Biostatistics, the application of statistics in biology-related fields and the limitation of such applications.

STATISTICS AND BIOSTATISTICS

The word statistics is used in two senses. It refers to collections of quantitative information, and to methods of handling that sort of data i.e. descriptive statistics. It also refers to the drawing of inferences about large groups on the basis of observations made on smaller one i.e. inferential statistics. Statistics, then, is to do with ways of collecting, organizing, summarizing and describing quantifiable data, and methods of drawing inferences and generalizing upon them. While the term Biostatistics is used when the data that are being analysed using statistical tools, are derived from the fields of biological sciences: Medicine, Pharmacy, Biochemistry, Microbiology, Agricultural Sciences and other biology-related areas.

USE OF STATISTICS IN BIOLOGY, AGRICULTURE AND MEDICINE.

Unlike other fields of science such as the physical sciences of chemistry and physics, variation is regarded as a fundamental feature in natural sciences of biology, agriculture and medicine. Biostatistics helps to explain this natural variation inherent in these fields of natural sciences. For example, variation may occur due to age of the population or may occur among individuals of a population due to diseases or their genetic makeup. Experimental design is an important aspect of biostatistics that describe on how to collect, organize, summarize and analyze data such that valid and objective conclusions or decision about the population can be drawn. Therefore before applying statistics in research, a research must know:

- What technique to use for an investigation
- What to be achieved
- Rules of using the technique, using correctly the statistical techniques for analysis of biological data.
- Statistical significance test for comparing one set of data with another.
- Determination of relationship between two variables either the use of correlation or fitting the best straight line or curve on a graph

DISCRETE AND CONTINUOUS VARIABLES

To gain knowledge about secondly haphazard events, statistician collect information for variables which describe the event. Therefore, a variable is a characteristics attribute that can assume different value. Variables can be classified into two broad categories.

1. Qualitative variables
2. Quantitative variables

Qualitative variables are variables that can be placed into distinct categories, according to some characteristic or attribute. For example, if subjects are classified according to gender (i.e. male or female), then the variable gender is qualitative.

Quantitative variables are numerical and can be ordered or ranked. For example, the variable age is numerical, and people can be ranked according to the value of their ages. Quantitative variables can be grouped into two:

1. **Discrete Variables** – can be assigned values such as 0,1,2,3 (integers) and are said to be variables that assume values that can be counted. Examples include number of children in a family, number of birds in a pen, number of trees in a garden, number of animals per litter etc.
2. **Continuous variables** – can assume all values between any specific values. They are obtained by measuring. This applies to variables such as length, weight, height, yield, temperature and time, that can be thought of as capable of assuming any value in some interval of values.

SAMPLING

When a set of observations is collected from a population, the population mean (μ), population variance (σ^2) and population standard deviation (σ) can be computed from it as the properties of the population. In the case of a sample, the parameters that describe it are the sample mean (x), sample variance (s^2) and sample standard deviation (s). Since the sample is a portion of the population, the parameters of the sample represent an estimate of the true parameters of the population. Therefore, sampling is a random process of selecting a sample from a population selected for study.

SAMPLE: A sample is a subgroup of the population selected for study. When a sample is chosen at random from a population, it is said to be an unbiased sample. That is, the sample for the most part, is representative of the population. But if a sample is selected incorrectly, it may be a biased sample when some type of systematic error has been made in the selection of the subjects. However, the sample must be random in order to make valid inferences about the population.

IMPORTANCE OF SAMPLE / SAMPLING

A sample is used to get information about a population for several reasons:

1. It saves the researcher time and money.
2. It enables the researcher to get information that he or she might not be able to obtain otherwise.
3. It enables the researcher to get more detailed information about a particular subject

SAMPLING METHODS

In order to obtain unbiased samples, several sampling methods have been developed. The most common methods are random, systematic, stratified, and clustered sampling.

RANDOM SAMPLING

For a sample to be a random sample, every member of the population must have an equal chance of being selected. Therefore, a random sample is one that has the same chance as any other of being selected. Randomness assists in avoiding various forms of conscious and unconscious bias and can be achieved by these two ways:

1. Number each element of the population and then place the numbers on cards. Place the cards in a hat or bowl, mix them, and then select the sample by drawing the cards. You must ensure that the numbers are well mixed.
2. The second and most preferred way of selecting a random sample is to use random numbers e.g. Table of random numbers by Fisher and Yates. The table comprises of a series of digits 0, 1, 2…. up to 9 arranged as such that each number had the same chance of appearing in any given position.

STRATIFIED SAMPLING

A stratified sample is a sample obtained by dividing the population into subgroups, called strata, according to various homogenous (alike) characteristics and then selecting members from each stratum for the sample. For example, you can group the items on basis of their age, size, colour etc. The advantage of stratified sampling is that it increases precision because all types of groups are represented through stratification and a heterogeneous population is made into a homogenous one.

CLUSTER SAMPLING

A cluster sample is a sample obtained by selecting a preexisting or natural group, called a Cluster and using the members in the cluster for the sample. For example a habitat, or a large area or field is divided into smaller units and a number of such units are randomly selected and used as a sample. There are three advantages to using a cluster sample instead of other types of sample:

1. A cluster sample can reduce cost
2. It can simplify field work.
3. It is convenient. The major disadvantage of cluster sampling is that the elements in a cluster may not have the same variations in characteristics as elements selected individually from a population.

SYSTEMATIC OR SKIP SAMPLING

This method involves taking an item as a sample from a larger population at regular intervals. For example, when sampling from a poultry farm, every third or fifth or tenth chick coming out of the cage is taken and included in the sample. This is done after the first number is selected at random for counting to start.

PROPORTIONATE SAMPLING

This type of sampling involves selecting a sample in proportion to the different groups in the population under study. For example: Assuming in a given terrestrial habitat there are the following different proportions of organism:

Trees - 100
Shrubs - 150
Vertebrates – 60
Invertebrates - 250

In sampling such a population, you may wish to pick 15 trees, 10 shrubs, 5 vertebrates, 20 invertebrates. These will form your sample.

SAMPLING DISTRIBUTION

If a sample of n observation is taken at random from a population, the sample is expected to have a mean x. Suppose another sample also of n observations is taken from the same population, it will similarly have a mean x (as the first one). The numerical values of these means will differ slightly because even though the samples are taken from the same population, the representative members of the two samples differ. As you take samples from the same population so will you get different numerical values for their means. This set of numerical values is called the Sampling distribution of the mean and it is determined by the nature of the population and the sample size. Therefore, a sampling distribution is the distribution of values from a mass of samples, one value per sample. If the sample size is large then the sampling distribution of the mean approximates very closely to a normal curve. This implies that the mean of the sampling distribution of means is equal to the population mean.

Exercise.

Identify the following as either discrete or continuous data: (a) Number of patients coming to the hospital each day (b) Lifetimes of micro organism in a certain pond (c) Heights of 200 birds in a certain farm (d) Yearly salary of an agronomist (e) Temperatures recorded every 30 minutes of a little boy suffering from malaria fever.

Identify each of the following sampling methods. (a) Every fish in a fish pond has equal chances of being included in a sample (b) Every corn plant that is 2m apart in a particular farm will be included in a sample (c) In surveying piggeries within a region, we choose to select 50 pig farms blocks and then investigate every pigs within the selected blocks

SECTION TWO

FREQUENCY DISTRIBUTION

Measurements or counting gives rise to raw data. Raw data itself is difficult to comprehend because it lacks organization, summarization, which renders it meaningless. Thus, the raw data has to be put in some order through classification and tabulation so as to reduce its volume and heterogeneity. To describe situations, draw conclusions or make inferences about events, the researcher must organize the data in some meaningful way. The most convenient method of organizing data is to construct a frequency distribution.

The Frequency distribution

Frequency is the number of occurrences of an element in a sample and is symbolized by f. A frequency distribution is the organization of raw data in table form, using classes and frequencies. When data are collected in original form, that is as observed or recorded they are called raw data.

Types of Frequency Distribution

Two types of frequency distributions that are most often used are the:

Categorical Frequency:

This is used for data that can be placed in specific categories, such as nominal or ordinal-level data. It is useful to know the proportion of values that fall within a group, category or observation rather than the number of values or frequencies. To get the relative frequency, the frequency of occurrence of each number is divided by the total number of values and multiplied by hundred. This can be expressed as follows: $\frac{f}{n}\ x\ 100\%$ Where f = Frequency of the category class and n = total number of values.

For example: The data below represents the blood groups of 40 students in a Biostatistics class. Construct a frequency distribution for the data.

A AB B O O A B AB A B O O O A AB B B A O AB A O O A
AB B B A A B AB A O B AB O A B A B

UNGROUPED FREQUENCY DISTRIBUTION

This is a list of the figures in array form, occurring in the raw data, together with the frequency of each figure, i.e. a frequency is constructed for a data based on a single data values for each class.

For example: Given below, are the wing length measurements (to the nearest whole millimeter) of 50 laughing doves.

76 73 75 73 74 74 72 75 76 73 68 72 78 74 75 72 76 76 77 70 78 72 70 74 76 75 75 79 75 74 75 70 73 75 70 74 76 74 75 74 78 74 75 74 73 74 71 72 71 79

GROUPED FREQUENCY DISTRIBUTION

The heights in inches of commonly grown herbs are shown below. Organize the data into a frequency distribution with six classes, and make useful suggestions. 18 20 18 18 24 10 15 12 29 36 13 20 18 24 18 16 16 20 7

SECTION THREE

PROBABILITY DISTRIBUTIONS

INTRODUCTION

Probability is a branch of mathematics which as a general concept can be defined as the chance of an event occurring. It is the basis of inferential statistics. This unit looks at three particular distributions, the Normal, the Poisson and the Binomial – all of which are important in sampling theory.

PROBABILITY DISTRIBUTION

A distribution is a scatter of related values, such as the assortment of weights in a group of cattle. A frequency distribution shows us how many times given values in a range of values occur. A Probability distribution is very similar because it shows us how probable given random variable values in a range of such values are. For example: if we toss two coins we can obtain 0, 1 or 2 'heads'. If we prepare a table showing the probabilities of all the random variable values we will have the probability distribution as shown below. (get the table as exercise).

THE NORMAL DISTRIBUTION

This is the most important distribution in statistics. It is also known as the Gaussian distribution named after Gauss, a German astronomer who showed its use in statistics. The normal distribution is defined by just two statistics, the mean and the standard deviation. Normal distribution is concerned with results obtained by taking measurements on continuous random variable (i.e the quantified value of a random event) like weight, yield etc. The normal distribution is a particular pattern of variation of numbers

around the mean. It is symmetrical (hence we express the standard deviation as ±) and the frequency of individual numbers falls off equally away from the mean in both directions. In terms of human height, progressively larger and smaller people than the average occur symmetrically with decreasing frequency towards respectively giants or dwarfs. What is important about this distribution is not only that this kind of natural variation often occurs, but also that it is the distribution which comes with the best statistical reference for data analysis and testing of hypotheses. It so happens that the curve given by this probabilities distribution approximates very closely to a Mathematical curve. This curve is called the Normal curve. In checking for normality, it is important to know whether an experimental data is an approximate fit to a normal distribution. This is easily checked with large samples. There should be roughly equal numbers of observations on either side of the mean. Things are more difficult when we have only a few samples. In experiments, it is not uncommon to have no more than three data per treatment. However, even here we can get clues. If the distribution is normal, there should be no relationship between the magnitude of the mean and its standard deviation.

Properties of a Normal Curve

- It is a Unimodal symmetrical curve
- The mean, mode & median all coincide, thereby dividing the curve into two equal parts
- Most items on the curve are clustered around the mean
- No kurtosis or skewness in the curve
- The area beneath the curve is proportional to the observation associate with the part

POISSON DISTRIBUTION

A Poisson distribution is a discrete probability distribution that is useful when n is larger and p is small and when the independent variables occur over a period of time. It can be used when a density of items is distributed over a given area or volume, such as the number of plants growing per acre. It can also be used to discover whether organisms are randomly distributed. For example, in ecological studies, Poisson distribution is used to describe the spread of organisms like insects, trees, and snails' etc. by the following:

1. Divide the large area into small squares of equal size
2. Count the particular animal or plant species under study in each square
3. You can also randomly select a number of squares, if the area is two large. The probability of X occurrences in an interval of time, volume, area etc. for a variable where λ (lambda) is the mean number of occurrences per unit (time, volume, area etc) is given by:……find out.

For Example: In a study on the distribution of tree-roosting birds, if there are 200 birds randomly distributed on 500 trees, find the probability that a given tree contains exactly three birds.

BINOMIAL DISTRIBUTION

A binomial experiment is a probability experiment that satisfies the following four requirements:

1. Each trial can have only two outcomes or outcomes that can be reduced to two outcomes. i.e these outcomes can either be success or failure. No two events can occur simultaneously.
2. There must be a fixed number of trails.
3. The outcomes of each trial must be independent of each other.

4. The probability of a success must remain the same for each trial. A binomial distribution is a special probability distribution that describes the distribution of probabilities when there are only two possible outcomes for each trial of an experiment.
Examples: 1. The answer to a multiple choice question (even though there are four or five answer choices) can be classified as correct or incorrect. 2. When tossing a coin, you get either a head or a tail.

4. In selecting individuals from human population you select either a male or a female, a boy or a girl etc. The binomial probability formula is given by:…….find out.

Where: P = Numerical probability of a success = P(s) q = Numerical probability of a failure = P(F) n = Number of trials x = The number of successes in n trials ! = Mathematical symbol called 'factorial'. So n! means multiple all the numbers in a count down from the total number in the sample. For example: 7! =7 x 6 x 5 x 4 x 3 x 2 x 1, and 4! = 4 x 3 x 2 x 1.

Exercise: 1. A survey on birds showed that one out of five fire finch was trapped, using mist net, in a given season. If 10 birds are selected at random, find the probability that 3 of the birds were trapped in the previous season.