AGR 302 AGRICULTURAL STATISTICS AND DATA PROCESSING

NATIONAL OPEN UNIVERSITY OF NIGERIA
FACULTY OF AGRICULTURAL SCIENCE

COURSE CODE: AGR 302

COURSE TITLE: Agricultural Statistics and Data Processing

AGR 302: AGRICULTURAL STATISTICS AND DATA PROCESSING

COURSE GUIDE

NATIONAL OPEN UNIVERSITY OF NIGERIA

Course Code: AGR 302
Course Title: Agricultural Statistics and Data Processing
Course Writer: Professor Isaac A. Adeyinka
Ahmadu Bello University, Zaria

Content Editor:

Course Coordinator:

NATIONAL OPEN UNIVERSITY OF NIGERIA
Headquarters
14/16 Ahmadu Bello Way
Victoria Island
Lagos
Abuja Annex
245 Samuel Adesujo Ademulegun Street
Central Business District
Opposite Arewa Suites
Abuja
e-mail: centralinfo@nou.edu.ng
URL www.nou.edu.ng

Introduction
Agricultural Statistics and Data Processing (AGR 302) is a second semester course. It is a two credit unit compulsory course which all students offering Bachelor of Science (BSc) in Agriculture must take.

Statistics is a familiar and accepted part of modern world that is concern with obtaining an insight into the real word by means of the analysis of numerical relationships. It is used in almost all fields of human endeavour.

Since this course Agricultural Statistics and Data Processing entails analysis of numerical relationships, we will focus on the meaning of statistics and biostatistics (collections of quantitative information and method of handling such data, drawing inferences on the basis of observation). We will also discuss frequency of distribution, probability, hypothesis, correlation and regression, covariance, Analysis of Variance (ANOVA).

What you will learn in this course
In this course, you have the course units and a course guide. The course guide will tell you briefly what the course is all about. It is a general overview of the course materials you will be using and how to use those materials. It also helps you to allocate the appropriate time to each unit so that you can successfully complete the course within the stipulated time limit.

The course guide also helps you to know how to go about your Tutor-Marked-Assignment which will form part of your overall assessment at the end of the course. Also, there will be tutorial classes that are related to this course, where you can interact with your facilitators and other students. Please I encourage you to attend these tutorial classes.

This course exposes you to data collection, management and analyses, the knowledge will be helpful during your project in data collection and analysis. It is indeed very interesting field of agriculture and biology.

Course Aims
This course aims to enable you to know/understand the use of different statistical and biostatistical analysis and packages for agricultural sciences and agricultural data interpretations and inference.

Course Objectives
To achieve the aim set above, there are objectives. Each unit has a set of objectives presented at the beginning of the unit. These objectives will give you what to concentrate and focus on while studying the unit and during your study to check your progress.
The comprehensive objectives of the Course are given below.
At the end of the course, you should be able to:

- ✓ Discuss the use of statistics in area of agriculture
- ✓ Discuss the different sampling methods and understand the purpose and importance of sampling
- ✓ Mention type of frequency distribution.
- ✓ Organise data using frequency distribution.
- ✓ Explain the normal, poisson and binomial distributions

- ✓ Compute the probabilities in poisson and binomial probability distributions
- ✓ State the null and alternative statistical hypothesis
- ✓ Determine the level of confidence in a biological data
- ✓ Explain the relationship between type I and II errors
- ✓ Explain the purpose of goodness of fit test
- ✓ Compute correlation and regression
- ✓ Explain types of correlation and Regression
- ✓ Explain the principle of experimental design
- ✓ Define ANOVA and test statistical hypothesis using ANOVA
- ✓ Compute the simple Spearman correlation coefficient
- ✓ Give the difference between non-parametric and parametric test

Working through the Course
To successfully complete this course. you are required to read each study unit, read the textbooks and other materials provided by the National Open University.
Reading the reference materials can also be of great assistance.

There will be a final examination at the end of the course. The course should take you about17 weeks to complete.

This course guide provides you with all the components of the course, how to go about studying and how you should allocate your time to each unit so as to finish on time and successfully.

The Course Materials
The main components of the course are:

1. The Study Guide
2. Study Units
3. Reference/ Further Readings
4. Assignments
5. Presentation Schedule

Study Units
The study units in this course are given below:
AGR 302 Agricultural Statistics and Data Processing
Unit 1 Population and Sample
Unit 2 Frequency distribution, measures of location and measures of variation
Unit 3 Probability
Unit 4 Probability Distributions
Unit 5 Descriptive Statistics
Unit 6 Sampling, data collection and data processing techniques
Unit 7 Inference and hypothesis testing; Type I and type II errors
Unit 8 Analysis of Variance
Unit 9 Correlation and regression analysis
Unit 10 Analysis of Covariance
Unit 11 Hypothesis testing of attributes data
Unit 12 Goodness of fit
Unit 13 Chi-Square Test for Independence
Unit 14 Field experimentation, collection and processing of data

Each unit will take a week or two lectures, will include an introduction, objectives, reading materials, self assessment question(s), conclusion, summary, tutor-marked assignments (TMAs), references and other reading resources.

There are activities related to the lecture in each unit which will help your progress and comprehension of the unit. You are required to work on these exercises which together with the TMAs will enable you to achieve the objective of each unit.

Presentation Schedule
There is a time-table prepared for the early and timely completion and submissions of your TMAs as well as attending the tutorial classes. You are required to submit all your assignments by the stipulated date and time. Avoid falling behind the schedule time.

Assessment
There are three aspects to the assessment of this course.
The first one is the self-assessment exercises. The second is the tutor-marked assignments and the third is the written examination or the examination to be taken at the end of the course.
Do the exercises or activities in the unit applying the information and knowledge you acquired during the course. The tutor-marked assignments must be submitted to your facilitator for formal assessment in accordance with the deadlines stated in the presentation schedule and the assignment file.
The work submitted to your tutor for assessment will account for 30% of your total work.
At the end of this course you have to sit for a final or end of course examination of about a three hour duration which will account for 70% of your total course mark.

Tutor Marked Assignment
This is the continuous assessment component of this course and it accounts for 30% of the total score. You will be given four (4) TMAs by your facilitator to answer. Three of which must be answered before you are allowed to sit for the end of the course examination.
These answered assignments must be returned to your facilitator.
You are expected to complete the assignments by using the information and material in your reading references and study units.
Reading and researching into the references will give you a wider view point and give you a deeper understanding of the subject.
1 Make sure that each assignment reaches your facilitator on or before the deadline given in the presentation schedule and assignment file. If for any reason you are not able to complete your assignment, make sure you contact your facilitator before the assignment is due to discuss the possibility of an extension. Request for extension will not be granted after the due date unless there is an exceptional circumstance.
2 Make sure you revise the whole course content before sitting for examination. The self-assessment activities and TMAs will be useful for this purposes and if you have any comments please do before the examination. The end of course examination covers information from all parts of the course.

Course Marking Scheme	**Marks**
Assignment	
Assignment 1-4	Four assignments, best three marks of the four count at 10% each - 30% of course marks.
End of course examination	70% of overall course marks
Total	100% of course materials

AGR 302 Agricultural Statistics and Data Processing
Course Title Agricultural Statistics and Data Processing
Course Writer Professor I.A. AdeyinkaDR ILIYA S. NDAMS
Ahmadu Bello University, Zaria
Content Editor:

Course Coordinator:

National Open University of Nigeria
Headquarters
14/16 Ahmadu Bello Way
Victoria Island
Lagos
Abuja Annex
245 Samuel Adesujo Ademulegun Street
Central Business District
Opposite Arewa Suites
Abuja
e-mail: centralinfo@nou.edu.ng
URL www.nou.edu.ng

MODULE 1. Basic concepts of statistics

Unit 1. Population and Sample

1.0 Introduction

To many students, statistics is not different from mathematics. In this unit you will be introduced to the basic concept of statistics. Two basic concepts of statistics are population and sample. You will appreciate the difference between population and sample. You will also learn the terminologies associated with population and those associated with sample.

1.1 Objectives

At the end of this unit, you should be able to:

i) Define the population and sample

ii) Define a sample and sampling

iii) Discuss discrete and continuous variables

iv) Discuss the different sampling methods and understand the purpose and importance of sampling and the advantages made possible by sampling

1.3 Main Content

1.3.1 Definition of Population and sample as basic concept of statistics

1.3.1.1 Population

Definition: Population is the collection of all individuals or items under consideration in a statistical study (Weiss, 1999). The Population is the whole set of values or individuals you are interested in. The population may also be defined as the set of entities under study. An example is the weight of cattle in Nigeria. The cattle population include all bulls and cows currently alive, those that had lived and now dead and the ones that will live in the future. You will not be able to measure the weights of the entire cattle population because many cattle are yet unborn while

many are already dead and unreachable. Even when it is possible to reach all of them, it is often too costly in terms of money and time involved. In the example you are interested in the population of cattle and your parameter of interest is body weight.

1.3.1.2 Sample

Sample is the part of the population from which information is collected (Weiss, 1999)

Since you cannot reach all the members of the population to take measurement, you will take a subset of population. This subset is called sample. You will then use this subset to draw inferences about the population under study, given some conditions. You will therefore take a subset of cattle population which is called sample, measure their weights and calculate the average or mean. The means that you calculated from sample is called a statistic. It is this statistic that you will use to draw an inference about the parameter of the population of interest. Because of the uncertainty and inaccuracy involved in drawing conclusions about the population based upon sample, you can only draw an inference about the population. You should take note that you will always have few numbers in your sample than the population. So you are bound to lose some information on the population.

1.3.2 Sample and Sampling

You must ensure that your samples are randomly selected to avoid bias in the use of the statistic to estimate the parameter. This is done using the Simple Random Sampling. In Simple Random Sampling each member of the population has an equal probability of being included in the sample. That is why it is called random.

1.3.3 Discrete and continuous variables

A variable is any characteristic or trait that varies or changes when moving from an individual to individual or object to object in a collection. If you conduct an study or an experiment, several

variable are involved. You may be interested in the sex of the animal, the parity of the animal, number of tits of the udder of the animal. When counting rather than measuring is involved, the result is a discrete random variable. Discrete random variable can only take on a certain number of integer value within an interval.

Example of discrete random variable would be number of kids born by goat, number of piglets farrowed by sows etc.

Continuous random variables are the result of a measurement on a continuous number scale. Example is birth weight goat.

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 animals 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 piglets farrowed by a sow, 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.

1.4 Tutor marked Assignment

1. Define population in your own words

2. Define sample also in your own word.

3. What is the relationship between population and sample

4. Define variable. Can you differentiate between discrete and continuous variable

5. Classify the following variable

a) Weight of broiler chicken at 8 weeks

b) Sex of day hold pullet chicks

c) Pollness in cattle (whether a cow or bull has or does not have horn)

d) Number of egg produced by chickens

e) Body length of goat

f) Number of parity of a cow or goat or sheep

g) Litter size

h) Litter weight

i) Staff strength on the poultry farm

j) Milk yield

1.5 References

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

UNIT 2: FREQUENCY DISTRIBUTION

2.1 Introduction

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.

2.2 OBJECTIVES:

At the end of this unit, you should be able to

i) Define frequency and frequency distribution.

ii) Mention the types of frequency distribution

iii) Organize data using frequency distributions

iv) Give reasons for constructing distribution.

v) Represent data in methods other than frequency distribution.

2.3 MAIN CONTENTS

2.3.1 The Frequency distribution

The organization of raw data in tabular form using classes (or intervals) and frequencies is known as 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.

2.3.2 Types of Frequency Distribution

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

i. 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. Examples are cattle coat color, pollness in cattle, sex of calves etc. 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:

$$= \frac{f}{n} \times 100\%$$

Where f = Frequency of the category class and n = total number of values.

Example

In a herd of cattle, 50 animals were tested for mastitis. The raw data is presented below with + representing animal with mastitis and – representing those without mastitis.

Animal ID	Mastitis Status	Animal ID	Mastitis Status	Animal ID	Mastitis Status	Animal ID	Mastitis Status	Animal ID	Mastitis Status
1	+	11	+	21	+	31	-		-
2	-	12	-	22	-	32	+		+
3	-	13	+	23	+	33	-		-
4	+	14	-	24	N	34	N		+
5	-	15	-	25	-	35	-		-
6	-	16	-	26	+	36	-		-
7	+	17	N	27	-	37	+		-
8	N	18	-	28	N	38	-		+
9	+	19	+	29	+	39	+		-
10	N	20	N	30	-	40	N		N

+ represented Mastitis
- represented no Mastitis
N – Not proved

Solution:

Since the data are categorical, the data can be grouped to Mastitis, No Mastitis and undecided as the classes for the distribution.

The following table was generated

Class	Tally	Frequency	Percent
Mastitis	~~////~~ ~~////~~ ~~////~~ //	17	34
No Mastitis	~~////~~ ~~////~~ ~~////~~ ~~////~~ ////	24	48
Undecided	~~////~~ ////	9	18

Therefore, you can conclude that in the sample more cows are without mastitis compared to those with mastitis because the frequency is the highest for those without mastitis.

2.3.3 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.

Example

You were asked to record the number of eggs produced by each of 50 birds when they attained 180 days of age

78	75	77	75	76	76	74	77	78	75
70	74	80	76	77	74	78	78	79	72
80	74	72	76	78	77	77	81	77	76
77	72	75	77	72	76	78	76	77	76
80	76	77	76	75	76	73	74	73	81

Construct the frequency distribution of the above data.

SOLUTION: The measurements above are presented in the order in which the observations were recorded. This can be represented in an ordered array so that the minimum and maximum values can easily be read

70	72	72	72	72	73	73	74	74	74
74	74	75	75	75	75	75	76	76	76
76	76	76	76	76	76	76	76	77	77
77	77	77	77	77	77	77	77	78	78
78	78	78	78	79	80	80	80	81	81

Find the range of the data: Highest value – lowest value (81 – 70 =11). Since the range of the data is small, classes of single data values can be used.

When you divide the frequency for any class by the total number of observation, what you get is called relative frequency

$$\text{relative frequency} = \frac{\text{frequency for class}}{\text{total number of observation}}$$

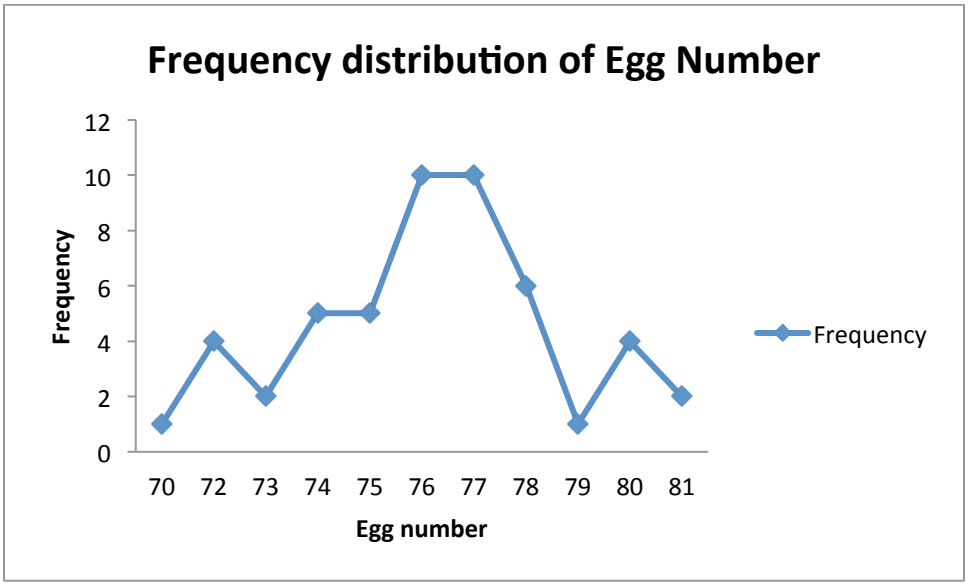
You can also express relative frequency in percentage.

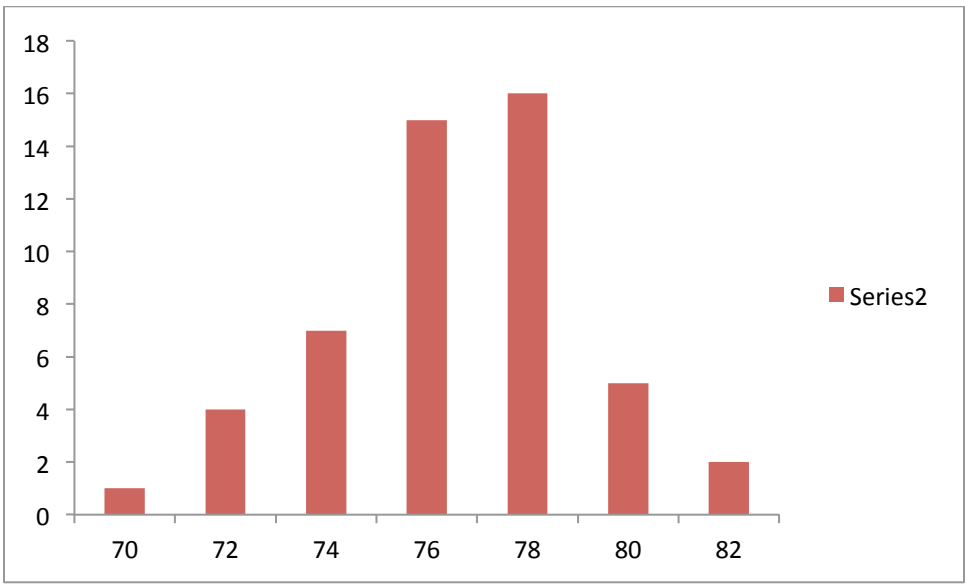
Cumulative frequency is the sum of the frequencies up to and including the given value.

Cumulative relative frequency for a specific value in a frequency table is the sum of the relative frequencies for all values at or below the given value.

Table 2.1: A tally of frequency of number of eggs produced by 50 chickens

Class limits	Tally	Frequency	Cumulative frequency	Relative frequency (%)
70	/	1	1	2
72	////	4	5	8
73	//	2	7	4
74	~~///~~	5	12	10
75	~~///~~	5	17	10
76	~~///~~, ~~///~~	10	27	20
77	~~///~~,~~///~~	10	37	20
78	~~///~~,/	6	43	12
79	/	1	44	2
80	////	4	48	8
81	//	2	50	4





2.3.4 Quantitative Frequency Distributions - Grouped

A grouped frequency distribution is obtained by constructing classes (or intervals) for the data, and then listing the corresponding number of values (frequency count) in each interval.

Rules to be followed in the construction of a frequency 1.

A frequency distribution should have a minimum of 5 classes and a maximum of 20 classes

The class width should be an odd number.

The class midpoint (Xm) is given by

$$Xm = \frac{\text{upper boundary+lower boundary}}{2} \text{ or } \frac{\text{upper limit+lower limit}}{2}$$

Midpoint is the numeric location of the center of the class

The classes must not have overlapping class limits e.g.

Incorrect class Overlapping	Correct Class Non Overlapping
0-10	0-10
10-20	11-20
20-30	21-30
30-40	31-40

The classes must be continuous, even if there are no values in a class ie there should be no gaps in the classes for lack of values.

Enough classes should be created to accommodate the whole data. i.e. every value in the data must belong to a class. Note: If zero frequency is the first or last, then it can be ignored.

The classes must be of equal width. In rule number 3 above, the class width is 10. Here, it is important to note that some times in open ended distribution i.e. distribution that has no specific beginning or ending value as:

Age	Temperature
11-20	5 and below
21-30	6-10
31-40	11-15
41-50	16-20
51 and above	21-25

In the class distribution above, any value above 51 is tallied in the last class while in the distribution for temperature, it simply means that any value below $5^{o}C$ will be tallied in the first class.

Effect of grouping

As a result of grouping, it is possible to detect a pattern in the figures but grouping results in the loss of information i.e. calculations made from a grouped frequency distribution can never be exact, and consequently excessive accuracy can only result in spurious accuracy

The reasons for constructing a frequency distribution are:

1. To organize the data in a meaningful, intelligible way.

2. To enable the reader to determine the nature and shape of the distribution.

3. To facilitate computational procedures for measures of average and spread.

4. To enable the researcher to draw charts and graphs for the presentation of data.

5. To enable the reader to make comparisons among different data sets.

Tutor Marked Assignment

1. Discuss the major reasons for constructing the frequency of data set
2. The following figures are the list of 20 day old chicks hatched from NAPRI Hatchery

   Classify the chicks according to their sexes and produce a frequency table

ChickID	1201	1202	1203	1204	1205	1206	1207	1208	1209	1210
sex	M	F	M	M	M	F	F	F	M	F

ChickID	1211	1212	1213	1214	1215	1216	1217	1218	1219	1220
sex	F	M	M	M	F	F	M	M	F	M

3. Using the table above, construct the relative frequency of the two sexes.
4. From your study, what should be the minimum and maximum number of classes in constructing group frequency? Support your explanation with examples.

References/Further reading

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

Daniel Navaro Learning Statistics with R: A tutorial for psychology students and other beginners (version 0.4). http://ua.edu.au/ccs/teaching/lsr

Unit 3 Probability

3.1 Introduction

Simply probability is the chance that something will happen. Mathematically, you can define probability as the extent to which an event is likely to occur, measured by the ratio of the favourable cases to the whole number of cases possible. It can also be defined as the measure of the likeliness that an event will occur. Every action that you take in your daily life is surrounded by the concept of probability. Example include weather forecast, football games etc.

3.2 Objectives

At the end of this unit, you will be able to

1. Use many terms widely used when talking about probability
2. Classify probability into classical, empirical and subjective probabilities
3. Understand basic properties of probability.

3.3 Main Content

3.3.1 Common terms you will often common across when talking about probability include the following:

Experiment: When you measure or observe an activity for the purpose of collecting data, the process is call experiment. E.g. rolling a pair of dice or tossing fair coin

Outcome: A particular result you obtained from an experiment is an outcome. For example, when you roll 2 dice, if you obtained double 6 or a pair of six, that is the out outcome of that experiment.

All of the possible outcomes of an experiment are referred to as **sample space**. When you roll 2 dice, the sample space of that experiment is the numbers {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}.

Any outcome or combinations of outcomes out of the sample space that are of interest is referred to as **event**.

3.3.2 Type of Probability

3.3.2.1 Classical Probability

When we know the number of possible outcomes of the event of interest and we also know the sample space, then we can calculate the classical probability of that event with the following equation:

$$P[A] = \frac{\text{Number of possible outcomes in which Event A occurs}}{\text{Total Number of possible outcomes in the sample space}}$$

Where:

P[A] = the probability that Event A will occur

A classical probability is the relative frequency of each event in the sample space when each event is equally likely.

Terms to note in the definition of classical probability are random, n, mutually exclusive, and equally likely.

Example:

If Event A=rolling 1 or 2 or 3 with one dice, you need to define the sample space for this experiment which is shown below:

{1} **{2}** **{3}**

{4} {5} {6}

There are 6 total outcome for this experiment. Each of the outcome has the same chance of occurring. The outcome that correspond to Event A are bolded. Therefore

$P[A] = \frac{3}{6} = 0.50$

Other examples are:

1. The roll of a die: There are 6 equally likely outcomes. The probability of each is 1/6.

2. Draw a card from a deck: There are 52 equally likely outcomes.

3. The roll of two die: There are 36 equally likely outcomes (6x6): 6 possibilities for the first die, and 6 for the second. The probability of each outcome is 1/36.

4. Drawing (with replacement) four balls from an urn with an equal number of red, white, and blue balls: There are 81 possible outcomes ($3x3x3x3 = 3^4$). For example, {red, white, white, blue} is an outcome which is a different outcome from {white, white, red, blue}. The probability associated with each outcome is 1/81.

5. The toss of two coins: The four possible outcomes are (H,H), (H,T), (T,H) and (TT). The probability of each is 1/4.

6. The draw of two cards: There are 522 possible outcomes.

A Basic assumption in the definition of classical probability is that n is a finite number; that is, there is only a finite number of possible outcomes. If there is an infinite number of possible outcomes, the probability of an outcome is not defined in the classical sense.

Definitions

Mutually exclusive: The random experiment result in the occurrence of only one of the n outcomes. E.g. if a coin is tossed, the result is a head or a tail, but not both. That is, the outcomes are defined so as to be mutually exclusive.

Equally likely: Each outcome of the random experiment has an equal chance of occurring.

Random experiment: A random experiment is a process leading to at least two possible outcomes with uncertainty as to which will occur.

Sample space: The collection of all possible outcomes of an experiment. If you had to guess, you would say it is called ”sample space” because it is the collection (set) of all possible samples.

An important thing to note is that classical probabilities can be deduced from knowledge of the sample space and the assumptions. Nothing has to be observed in terms of outcomes to deduce the probabilities.

3.3.2.2 Empirical probability

When you don’t know enough about the underlying process to determine the number of outcomes associated with an event, we rely on empirical probability. This type of probability observes the number of occurrences of an event through an experiment and calculates the probability from a relative frequency distribution. Therefore:

$$P[A] = \frac{\text{Frequency in which Event A occurs}}{\text{Total Number of observation}}$$

One example is probability that Musa will feed the chickens after his mother’s first shout?

The following table indicates the number of shouts Musa required over the last 20 days before he goes to feed the chickens.

Musa’s Shout from mum before feeding the chickens									
2	4	3	3	1	2	4	3	3	1
4	2	3	3	1	3	2	4	3	4

You can summarize the data in the following table

Number of mum's shout	Number of observation	Percentage
1	3	3/20 = 0.15
2	4	4/20 = 0.20
3	8	8/20 = 0.40
4	5	5/20 = 0.25
Total	20	

Based on these observation, if Event A= Musa getting up to feed the chickens on the first shout from his mother then P[A] = 0.15

Using the table above you can you can get the probability of other events. Let us say that Event B = Musa requiring more than 2 shouts to get the chickens fed then P[B] = 0.40 + 0.25 = 0.65.

If you run another set of 20 day experiment, you will get a different result than the previous one. If you run the experiment for 100 days, the relative frequency would approach the true or classical probabilities of the underlying process. The pattern is known as the law of large numbers.

3.3.2.3 Subjective probability

Subjective probability is used when classical and empirical probabilities are not available. Under this condition we rely on experience and intuition to estimate the probabilities of event A is the limit as n goes to infinity of (m/n) where m is the number of times that A is satisfied in the experiment and n is the number of times you run the experiment. Basically, it is the probability you will observe the event given that you run an infinite number of experiments like tossing a die an infinite number of times and seeing how often you get a "6".

Subjective probability is your belief of what the probability is. If someone believes that the probability of getting heads is 0.6 then this is a subjective probability.

Tutor marked assignment

1. Classify probability into 3 types and define each class or type of probability.
2. A student was given a list of 20 chickens. The tag number was from 1 – 20 of each chicken was written on small piece of paper and each one is rolled in to a ball. Each paper ball has an equal chance of being picked from a basket.
   a) What is the probability of chosen chicken number 1?
   b) What is the probability of chosen chickens number 3 and 10?
   c) What is the probability of chosen Chicken number 3 or 10?
3. You where given a fair die and asked to throw it once.
   a. What is the probability of throwing 1?
   b. What is sample space for this experiment?
   c. Construct a probability table for throwing 1, 2, 3, 4, 5 and 6?
   d. Can you construct a bar graph for the values?

3.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

UNIT 4 Probability Distributions

4.1 Introduction

In life, there is no certainty. Every event that comes our away is always associated with some level of uncertainty. When you plant some seeds of corns, there is some probability that this corn will germinate. The probability of the corn germinating may be improved if certain condition is fulfilled. For example, if the soil is well water, the probability of germination is improved. As you go through this unit, you will understand the concept of probability distribution as it is related to agricultural experiments and research.

4.2 Objective

You will be able to

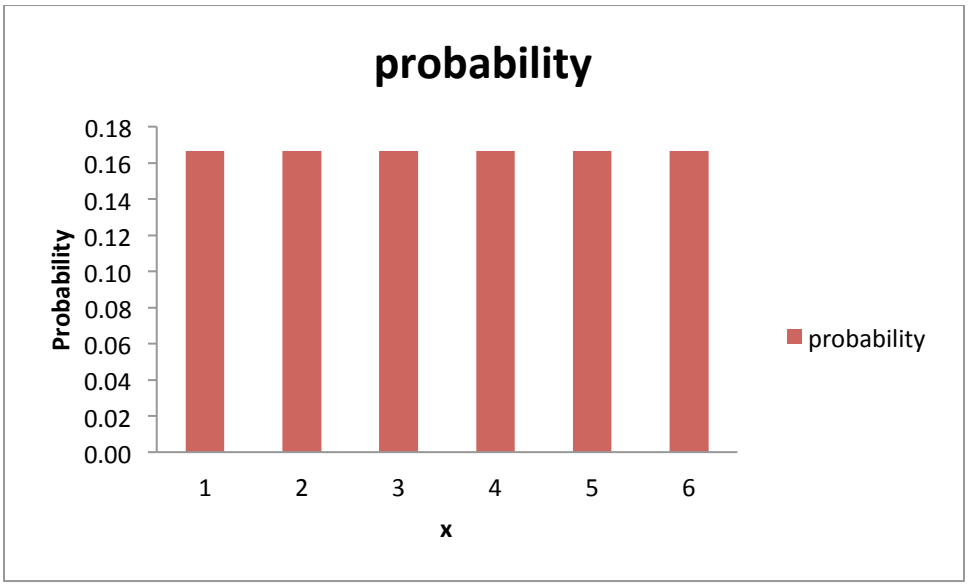
1. Define "distribution"
2. Interpret a frequency distribution
3. Distinguish between a frequency distribution and a probability distribution
4. Construct a grouped frequency distribution for a continuous variable
5. Identify the skew of a distribution
6. Identify bimodal, leptokurtic, and platykurtic distributions

4.3 Main Content

4.3.1 Probability distributions for Discrete Random variables

The probability distribution of a discrete random variable y is the table, graph or formula that assign the probability P(y) for each possible value of the variable y. A random variable is an outcome that takes on a numerical value as a result of an experiment. The value is not known with certainty before the experiment. But you know the sample space of the experiment. You

can denote the value of the random variable as x. For example in an experiment where a single dice is rolled, the P(x=1) = 1/6, P(x=2)=1/6, P(x=3) = 1/6, P(x=4)=1/6, P(x=5) = 1/6 and P(x=6)=1/6. The sum of all the probability is 1.

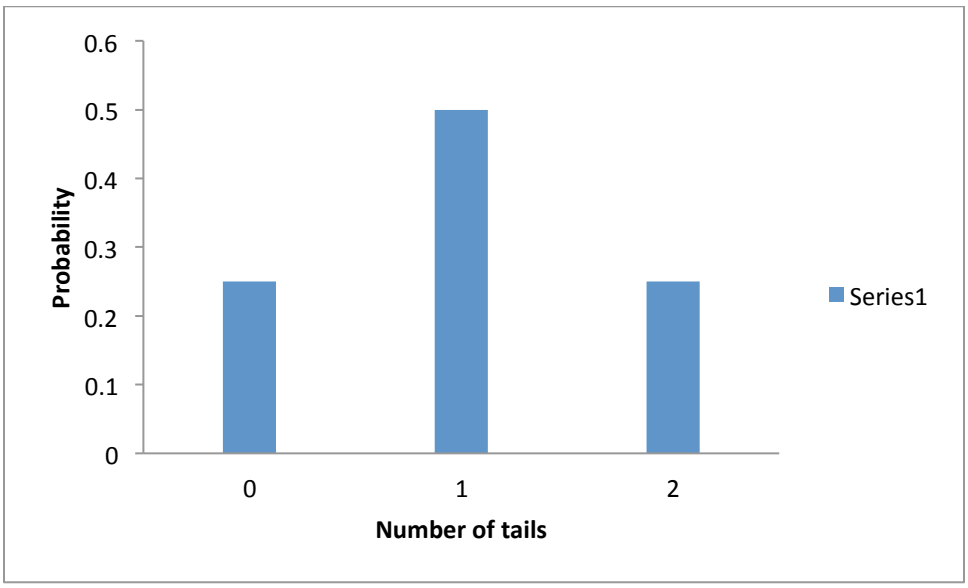


This is **a simple probability distribution**.

Another example is tossing two coins, you can obtain 0, 1 or 2 'tails'. You will then prepare a table showing the probabilities of all the random variable values.

Number of tails	Sequential event	Probability
0	HH	¼ = 0.25
1	HT, TH	¼ + ¼ = ½ =0.5
2	TT	¼ = 0.25
total		1.00

The sum of probability distribution is always equal to 1.



Above is shown the histogram of the probability distribution of 0 tail, 1 tail and 2 tails.

The rules of discrete probability distributions are as follows:

Each outcome in the distribution needs to be mutually exclusive – that is the value of the random variable cannot fall into more than 1 of the frequency distribution classes. For example you cannot have both head and tail in one toss of a coin.

The probability of each outcome P(x) must be between 0 and 1.

The sum of the probabilities for all the outcomes in the distribution needs to add up to 1.

4.3.2 Binomial Distribution

Binomial Distribution is the probability distribution that is associated with binomial experiment.

Binomial Experiment is an experiment with only 2 possible outcomes. For example the probability (p) of an animal dying before weaning is 0.2 while the probability (q) of not dying before weaning is 0.8.

Because only 2 outcomes are allowed in a binomial experiment, p =1 - q always hold true.

A binomial experience requires that each trial is independent of any other trials.

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.

The binomial probability distribution allows you to calculate the probability of a specific number of successes for a certain number of trials.

There must be a fixed number of trials.

Examples are

1. When you toss a coin, you can only get a head or a tail
2. When choosing people to participate in an experiment you can choose a male or a female
3. When you chose animals for a trial, an animal can either be a male or female.

You can calculate the probability of r success in n trials using the binomial distribution, as follows:

$$\mathrm{P[x]} = \frac{n!}{(n-x)!\,x!}\,\mathrm{p^x q^{n-x}}$$

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 × 6 × 5 × 4 × 3 × 2 × 1, and 4! = 4 × 3 × 2 × 1

For Example: 1. A survey on animals pest destroying crop in a village showed that one grass cutter out of five animal pest was trapped, using traps for catching destructive pest in a village, in a given season. If animal pests are selected at random, find the probability that 3 of the grasscutters were trapped in the previous season.

Solution:

n = 10, x = 3, p = 1/5 and q=1 – 1/5 = 4/5

Substituting the values

P[3] = $\frac{10!}{(10-3)!3!} \times 0.2^3 \times 0.8^{10-3}$

$$P[3] = \frac{10\times9\times8\times7\times6\times5\times4\times3\times2\times1!}{7\times6\times5\times4\times3\times2\times1\times3\times2\times1} \times 0.2^3 \times 0.8^{10-3}$$

$$P[3] = \frac{10\times9\times8}{3\times2\times1} \times 0.2^3 \times 0.8^{10-3}$$

=0.201

As the number of trials increase in a binomial experiment, calculating probabilities using the previous formula becomes cumbersome and time consuming. You should consider using the binomial probability table. The probability table is organized by values of n, the total number of trials. The number of success r, are the rows of each section whereas the probability of success, p, are columns. Notice that the sum of each block of probabilities for a particular value p adds to 1.0.

You can use Excel to calculate Binomial probabilities using BINOMDIST(r, n, p, cumulative)

Where

cumulative = FALSE if you want the probability of exactly r successes

cumulative = TRUE if you want the probability of r or fewer successes

Mean and standard deviation for the Binomial Distribution

You can calculate the mean of Binomial probability distribution using the following formula

$\mu = \text{np}$

Where

n = the number of trials

p = the probability

For our example of trapped grasscutters

$\mu = 10 \times 0.2$

= 2.0 grasscutters

You can calculate the standard deviation for a binomial probability distribution using the following formula

$\sigma = \sqrt{npq}$

where q = probability of a failure

for our example, the standard deviation for the distribution is as follows

$\sigma = \sqrt{npq} = x = \sqrt{(10)(0.2)(0.8)}$

=1.6

Normal Distribution

The normal distribution is the most important and most widely used distribution in statistics. Many biological variables followed a normal distribution. Normal distributions are bell shaped

graphically. The Y-axis in the normal distribution represents the "density of probability." Intuitively, it shows the chance of obtaining values near corresponding points on the X-axis.

Concept:

First, the area under the curve equals 1. Second, the probability of any exact value of X is 0. Finally, the area under the curve and bounded between two given points on the X-axis is the probability that a number chosen at random will fall between the two points.

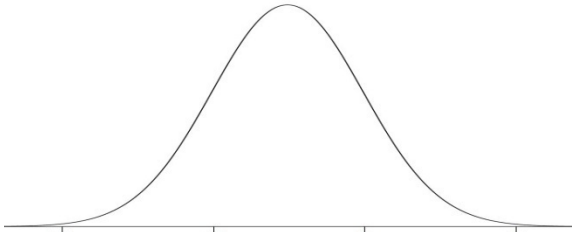
A normal distribution can be described by its means and its standard deviation. Normal distribution is concerned with results obtained by taking measurements on continuous random variable such (that is the quantified value of a random event) like height, length, age, weight, yield etc. Normal Distributions has some unique characteristics.

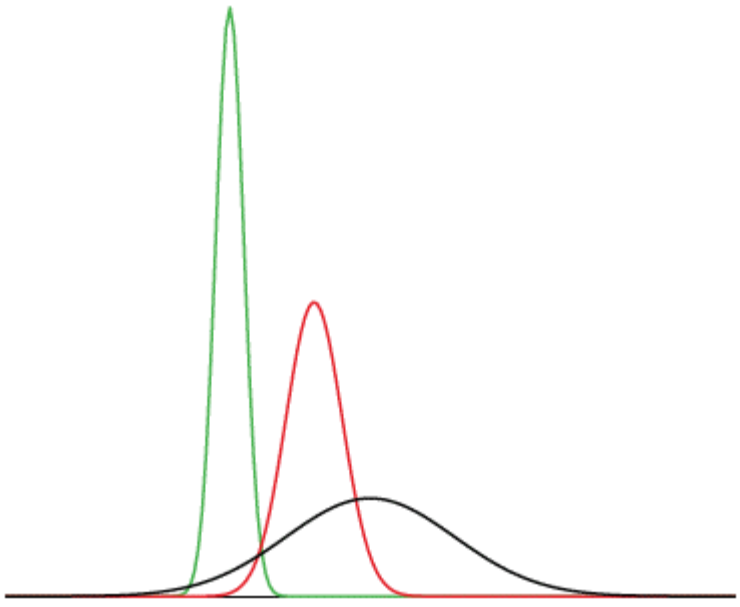
1. It has a kind of pattern of variation around the mean.
2. It is symmetrical hence you can say mean plus or minus standard deviation
3. the frequency of individual numbers falls off equally away from the mean in both directions

The normal distribution is extremely useful in statistics. Mathematicians proved that for samples that are “big enough,” values of their sample means,$\bar{x}$ (pronounced x-bar) (including sample proportions as a special case), are approximately distributed as normal, even if the samples are taken from really strangely shaped distributions. This important result is called the central limit theorem. In terms of body weights of cattle, progressively larger and smaller cattle than the average occur symmetrically with decreasing frequency towards respectively extremely large and extremely small cattle when large number of animals are weighed. The kind of variation observed with normal distribution is naturally occurring. The distribution also comes with the best statistical reference for data analysis and testing of hypotheses.

The curve given by this probabilities distribution approximates very closely to a *Mathematical curve* 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.



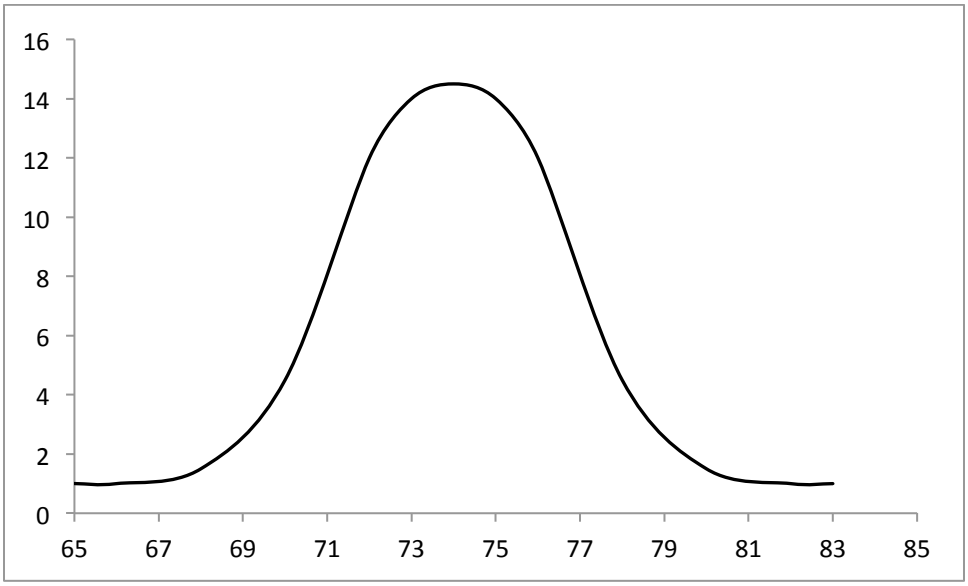


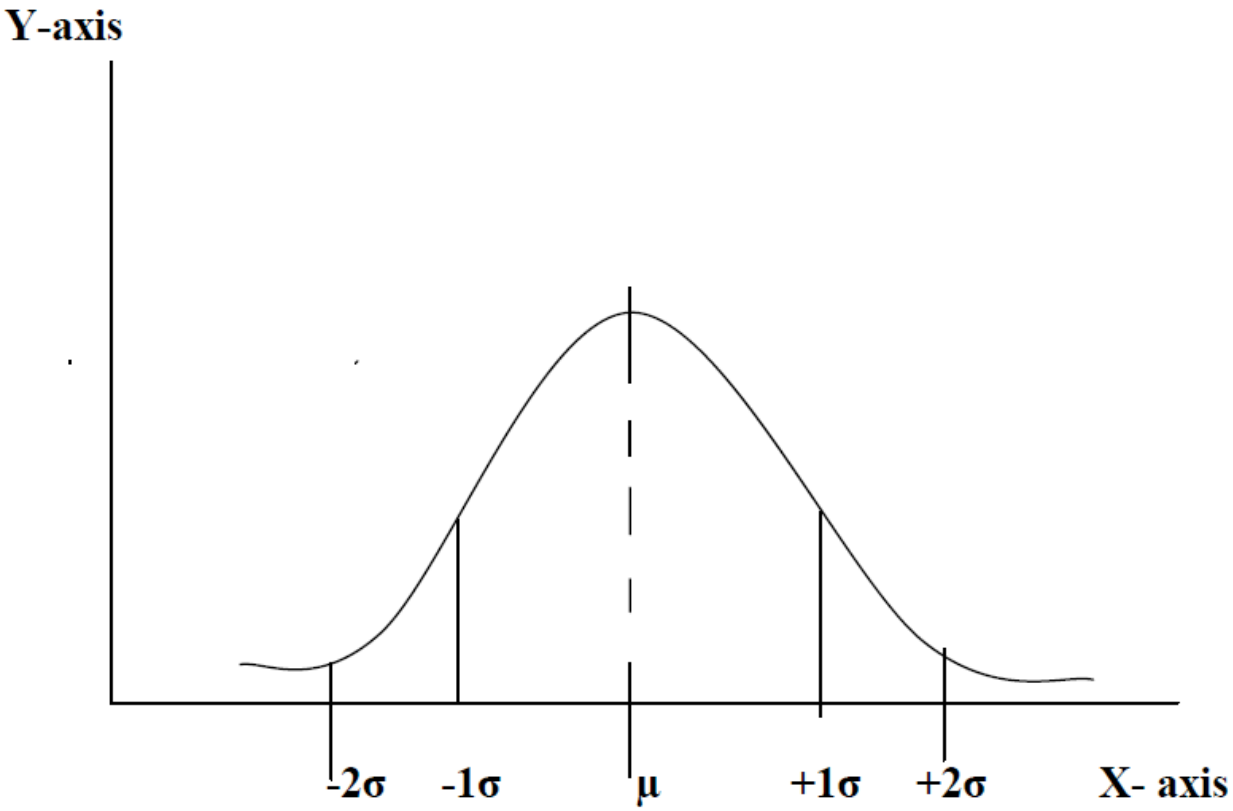
The green (left-most) distribution has a mean of -3 and a standard deviation of 0.5, the distribution in red (the middle distribution) has a mean of 0 and a standard deviation of 1, and the

distribution in black (right-most) has a mean of 2 and a standard deviation of 3. These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails.

The density of the normal distribution (the height for a given value on the x axis) is shown below. The parameters μ and σ are the mean and standard deviation, respectively, and define the normal distribution. The symbol *e* is the base of the natural logarithm and π is the constant pi.

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$





Normal curve with standard deviations

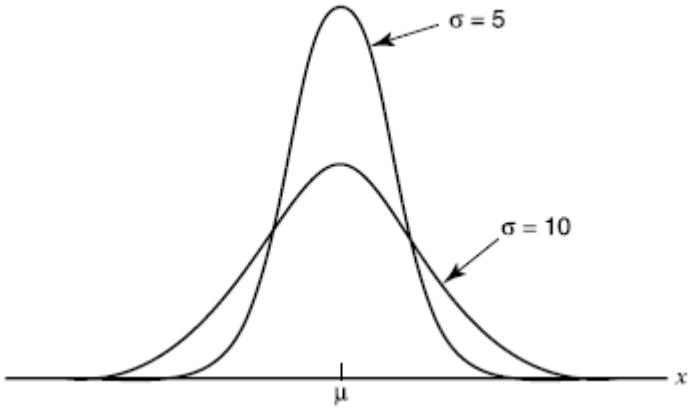
Properties of a Normal Curve

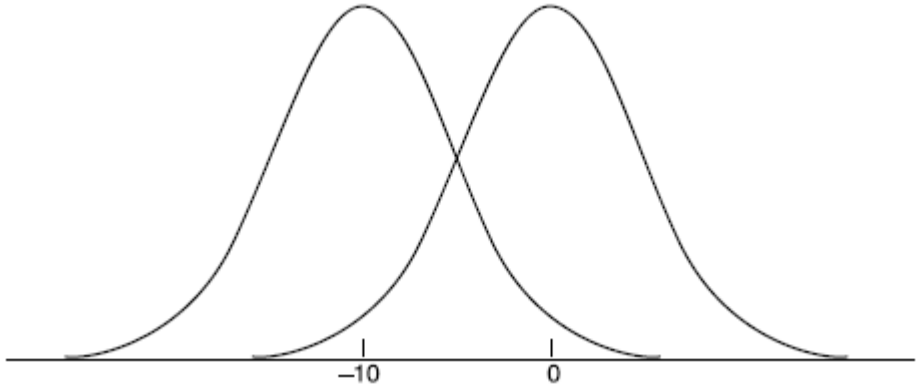
Eight features of normal distributions are listed below.

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
6. 68% of the area of a normal distribution is within one standard deviation of the mean.

7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.
8. Approximately 99% of the area of a normal distribution is within three standard deviations of the mean.

The term normal curve, in fact, refers not to one curve but to a family of curves, each characterized by a mean μ and a variance σ^2. In the special case where $\mu = 0$ and a variance $\sigma^2 = 1$, we have the standard normal curve. For a given μ and a given σ^2, the curve is bell-shaped with the tails dipping down to the baseline. In theory, the tails get closer and closer to the baseline but never touch it, proceeding to infinity in either direction. In practice, we ignore that and work within practical limits. The peak of the curve occurs at the mean m (which for this special distribution is also median and mode), and the height of the curve at the peak depends, inversely, on the variance σ^2. Some of these curves are shown in Figure 3.3.





Standardizing the Normal Distribution

Any value of an observation *X* on the baseline of a normal curve can be standardized as a number of standard deviation units, the observation is away from the population mean, μ. This is called a *z*-score. To transform *x* into *z* the formula is given by

$$z = \frac{(x - \mu)}{\sigma}$$

If the population mean μ is larger than the sample mean *x*, the *z* is negative.

If the sample size is more than about 30 observations, the sample mean (x) and standard deviation (s) are considered to be good estimates of μ and σ, and z is given by:

$$z = \frac{(x - \mu)}{s}$$

If the calculated value of z is larger than 1.96 (i.e. $P < 0.05$ or 95% confidence coefficient) then this is regarded as unlikely or statistically significant.

Tutor marked assignment

1. Differentiate between Binomial distribution and Poision Distribution. What are the characteristics of each of the distribution?

2. In a wild life conservation survey on birds, it was showed that one out of ten quails was trapped, using mist net, in a given season. If 20 birds are selected at

random, find the probability that 6 of the birds were trapped in the previous season.

3. A farmer has capacity for keeping just for goats.
    a. A farmer's goat is expecting a set of twin. What is the sample space for this farmer's expectation (hint: any combinations of the two sexes are possible)
    b. What is the probability the goat giving birth to at least 1 male?
    c. What is the probability of giving birth to 2 females.
    d. Construct a probability distribution for your answers.
4. What is Normal Distribution?
    a. What is the relationship between normal distribution, standard deviation and means?

4.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

Unit 4 Descriptive Statistics

5.1 Introduction

You will often be exposed to large amount of data that is difficult to understand without organizing them into fewer number that is easy to comprehend. In a way to understand your data, you must summarize the data. The methods for organizing and summarizing data to aid in effective presentation and increased understanding is the branch of statistics called **descriptive statistics.**

5.2 Objectives

After completing this Unit, you should be able to

1. Define "descriptive statistics"
2. Understand the concept of Univariate Analysis
3. Distinguish between descriptive statistics and inferential statistics

5.3 Main Content

5.3.1 Definition of Descriptive Statistics

Descriptive statistics can be defined as numbers that are used to summarize and describe data. The word "data" (singular "datum") means any information numerical logical or textual that have been collected through survey, designed experiment, observation of natural events or observation of historical records. Example of descriptive statistics is the total number of all cattle in Nigeria. In fact every livestock in Nigeria can be presented in Summary table showing total number of goats, sheep, cattle and chickens. We can also describe the cattle in Nigeria by their numbers, average weight, lactation length and total kg of milk per lactation. The chickens in Nigeria can also be described by origin either local or exotic, average body weight of local chickens and exotic chickens, Hen Day production of local and exotic chickens. Several

descriptive statistics are often used at one time to describe the data completely. Descriptive statistics does not involve generalizing to the population beyond the data that is available to you. The following table shows some descriptive statistics of ShikaBrown Chickens, a highly adapted egg type chickens

Rearing period livability (0-24 wks) %	90
Age at first egg, days	158
Age at 50% production, days	180
Age at peak production, days	210
Percent at peak production	75
Laying period livability (24-72 wks)%	82
Hen-housed egg production to 72 weeks, No	261
Hen-day egg production, No	270
Average hatchability, %	71.3
Body weight at maturity (40wks), kg	1.8
Average egg weight at maturity (40wks)	55.7
Body weight at maturity (72wks), kg	1.95

Descriptive statistics like the above give insight to the performance of ShikaBrown layer type chickens in Nigeria given a standard diet.

Let us look at another example.

The sexes of chicks that hatched out from NAPRI Hatchery over 10 hatches

Date Hatched	Pullet	Cockerels	Total
January	1240	1400	2640
February	4450	5150	9600
March	5100	5250	10350
April	4300	4100	8400
May	3500	3400	6900
June	6050	6010	12060
Total	24640	25310	49950

The above table summarized the number of chicks on a monthly basis from January to June. However you can see that there seems to be more male chicks (cockerels) than female chicks (pullet) every month. You may then ask a research question of what could be the cost of this difference or deviation from the expected sex ratio of male:female of 1:1? You may even imagine the possibility of tilting the sex ratio towards producing more females than males in a pullet chicks production enterprise as male chicks command a fraction of the value of the pullet chicks. Descriptive statistics help you to simplify large amounts of data in a sensible way. Descriptive statistics provide a powerful summary that may enable comparisons across animals or other units.

5.3.2 Univariate Analysis

Univariate analysis is the examination of data variable by variable. It is the simplest method of data analysis. There are three major properties of a single variable that you examine:

- the distribution
- the central tendency

- the dispersion

All of these properties will be used to describe each variable in any animal experiment.

The Distribution. This is a summary of the frequency of individual values or ranges of values for a trait or variable. The simplest distribution would list every value of a variable and the number of variates that had each value. For instance, a typical way to describe the distribution body weights of milking cows on a farm is by grouping the data on weight into a number of classes and listing the number or percent of cows that fell into each of the class or group. Or, you describe sex of calf born by listing the number or percent of males and females. In these cases, the variable has few enough values that we can list each one and summarize how many individual cows belong to each group or class. When you have large number of sample, you can group the data of a particular variable into a number of categories from a minimum of 5 to a maximum of 20.

Age Category	Percent
Under 35	9
36-45	21
46-55	45
56-65	19
66+	6

One of the most common ways to describe a single variable is with a frequency distribution. Depending on the particular variable, all of the data values may be represented, or you may group the values into categories first (e.g., with age, price, or temperature variables, it would usually not be sensible to determine the frequencies for each value. Rather, the value are grouped

into ranges and the frequencies determined). Frequency distributions can be depicted in two ways, as a table or as a graph. Table 1 shows an age frequency distribution with five categories of age ranges defined. The same frequency distribution can be depicted in a graph as shown in Figure 1. This type of graph is often referred to as a histogram or bar chart. Graphing is a way of visually presenting the data. Many people can grasp the information presented in a graph better than in a text format. The purpose of graphing is to:

- present the data
- summarize the data
- enhance textual descriptions
- describe and explore the data
- make comparisons easy
- avoid distortion
- provoke thought about the data

Bar Graphs

Bar graphs are used to display the frequency distributions for variables measured at the nominal and ordinal levels. Bar graphs use the same width for all the bars on the graph, and there is space between the bars. Label the parts of the graph, including the title, the left (Y) or vertical axis, the right (X) or horizontal axis, and the bar labels.

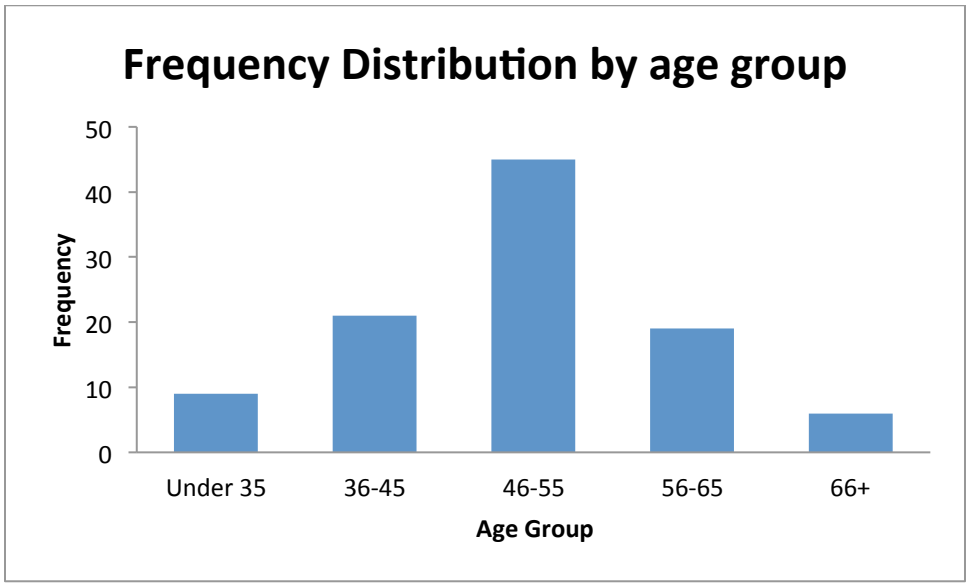
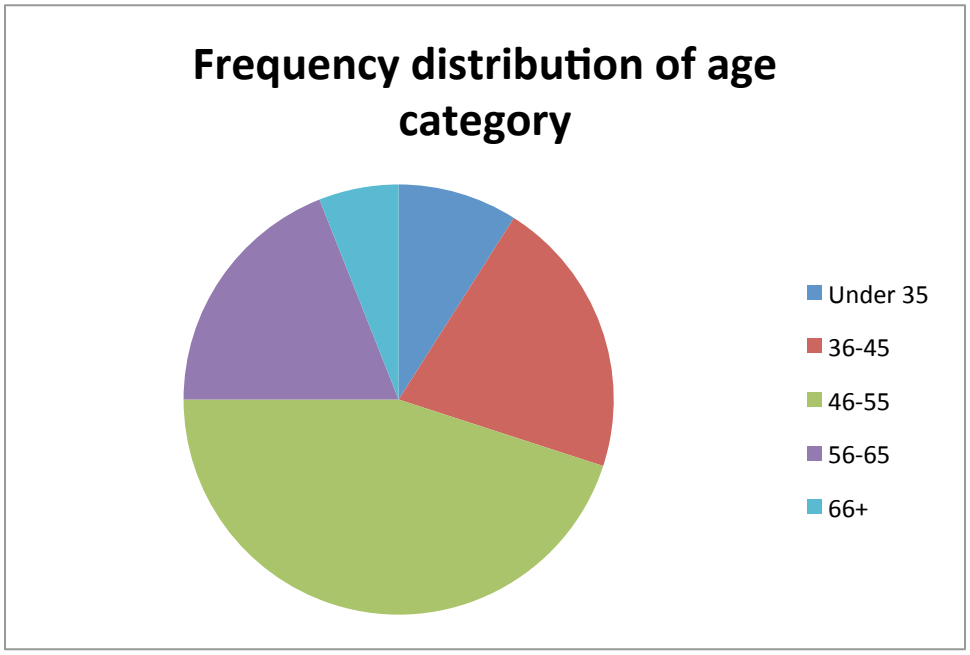


Figure 2. Frequency distribution bar chart.

Alternatively, you can also depict the same data using pie chart



Distributions may also be displayed using percentages. For example, you could use percentages to describe the:

- percentage of people in different income levels

- percentage of people in different age ranges
- percentage of people in different ranges of standardized test scores

Central Tendency. The central tendency of a distribution is an estimate of the "center" of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

The **Mean** or average is the most commonly used method of describing central tendency. To calculate the mean, add up all the values in your data and divide by the number of values. For example, you will determine the **mean** or **average** body weights of chicks at day old by adding up all the body weights at day old and dividing by the number of chicks that were weighed. For example, consider the weight of 10 chicks at day old.

35, 40, 41, 30, 36, 45, 35, 45, 35, 40

The sum of these 10 values is 382, so the mean is 382/10 = 38.2.

The **Median** is the score found at the exact middle of the set of values. To compute the median of a variable, you will list all scores in numerical order, and then find the score in the center of the sample. If you order the 8 scores shown above, you would get:

30 35 35 35 36 40 40 41 45 45

There are 10 weights and weights no 5 and 6 represent the halfway point. The weights are 36 and 40. So the median is

(36+40)/2 = 38

The **mode** is the most frequently occurring value in the set of body weights. You will determine the mode by the following procedure.

Arrange your data in order either ascending or descending.

Then count how many time each number occurred.

The most frequently occurring value is the mode.

Using your chick weight data, which of the data is the most frequently occurring? If your answer is 35 then congratulations you are correct!!

A frequency distribution with 2 modes is said to be bimodal.

Using the chick weight example you can form the following table

Parameter	Value
Mean	38.2
Median	38.0
Mode	35.0

Notice that for the same set of 10 weights we got three different values – 38.2, 38 and 35 for the mean, median and mode respectively. If the distribution is truly normal (i.e., bell-shaped), the mean, median and mode are all equal the same..

Now consider the weights of two groups of chickens at 2 weeks of age. Four chickens were taken from two different pens to see what they weigh at 2 weeks of age.

Chicken	Pen	
	A	B
1	100	134
2	150	137
3	80	129
4	200	130
5	90	138
6	160	140
7	90	125

8	190	127
9	125	135
10	135	125

What is the mean weight for chickens in Pen A? How does it compare with the mean of those that belong to Pen B?

Solution

Mean for Pen A Chickens = (100 + 150 + …+ 125+135)/10

= 132

Mean for Pen B Chickens = (134 + 137 + …+ 135+125)/10

= 132

You can also check for the median and mode of the two groups of chickens.

Your result should look as in the following table.

	Pen A	Pen B
Mean	132	132
Median	130	132
Mode	90	125

Dispersion

If you look at the 10 data points for each of the groups in the above exercise, you will quickly notice that the body weights of the chicks are not the same. But the means of the two group are the same. Within each of the group, some chicks weighed higher than the group mean while others weigh less. This means that each weight of the chick is dispersed around the common mean. In other words, each weight of the chicks can be written as mean plus or minus a quantity. There are measures for this dispersion that you should know.

Listed below are the common measures of dispersion

- Range
- Inter Quartile Range
- Mean absolute Deviation
- Variance
- Standard deviation
- Coefficient of variation

Dispersion refers to the spread of the values around the central tendency. There are two common measures of dispersion, the range and the standard deviation.

Range

The range is simply the highest value minus the lowest value. In the group A chickens, the minimum value is 80 and maximum weight is 200.

Range for group A is Maximum of A – Minimum of A

= 200 – 80 = 120

Range for group B is Maximum of B – Minimum of B

= 140 – 125 = 15

Obviously group A has a wider range than Group B.

Even though the range is easiest measure of dispersion to calculate, it is not considered a good measure of dispersion as it does not utilize the other information related to the spread. The Outliers, either the extreme low value or extreme high value can affect the range considerably.

Inter-Quartile Range

Inter-Quartile range is a measure of dispersion, which is not affected by the outliers of the data, but nevertheless conveys the idea of range. it measures the spread of the middle 50% of an ordered data set. In other words it is the numerical difference between the first and the third Quartiles.

Inter quartile Range = $Q_3 - Q_1$

Mean Absolute Deviation (MAD)

To calculate this parameter, the mean or average is subtracted from each of the value that contributed to the mean. The resulting number will either be negative or positive. However you will disregard the signs (you will take every value as positive), add them up and divide by number of individuals used in computing the mean.

For group A

MAD = $\frac{\sum|x-\bar{x}|}{N}$

(|100-132| + |150-132| + |80-132| + |200-132| + |90-132| + |160-132| + |90-132| + |190-132| + |125-132| + |135-132|)/10

=(|-32|+|18|+|-52|+|68|+|-42|+|28|+|-42|+|58|+|-7|+|3|)/10

$= \frac{|32|+|18|+|-52|+|68|+|-42|+|28|+|-42|+|58|+|-7|+|3|}{10}$

$= \frac{32+18+52+68+42+28+42+58+7+3}{10}$

= 35

And for group B

MAD for group B

(|2|+|5|+|-3|+|-2|+|6|+|8|+|-7|+|-5|+|3|+|-7)/10

= 4.8

Group A has a wider variation than Group B

So, what does this measure of dispersion tell about the spread? A data set with a larger Mean absolute difference is more spread when compared to a data set with a smaller MAD. The Mean absolute difference is sensitive to the outliers.

Standard Deviation

This the most commonly used estimate of dispersion. The Standard Deviation shows the relation that set of scores has to the mean of the sample. No consider the weights of ten chickens in group A

No calculate the means body weights of the ten chickens.

Then subtract the mean from each of the individual body weights. Then find the squares of each of the deviates. Add up the squares and divide your result by the number of observation less one.

The formula for calculating variance is

$$\sum(x-\bar{x})/N\text{-}1$$

You can easily do this in a table as follows

Sno	Weight	$x-\bar{x}$	$(x-\bar{x})^2$
1	100	-32	1024
2	150	18	324
3	80	-52	2704
4	200	68	4624
5	90	-42	1764
6	160	28	784
7	90	-42	1764
8	190	58	3364

9	125	-7	49
10	135	3	9
Mean	132		16410
Variance			1823.333
SD			42.70051

In order to calculate the standard deviation, you first find the distance between each value and the mean. Any value that is greater than the mean will result in a positive deviation. Those values that are below the mean will yield a negative deviation.

You can present the complete process of computation in a table as follows

Sno	**Day old chick weight (X)**	Mean ($\overline{X}$)	X-$\overline{X}$	(X-$\overline{X}$)^2
1	**35**	38.2	-3.2	10.24
2	**40**	38.2	1.8	3.24
3	**41**	38.2	2.8	7.84
4	**30**	38.2	-8.2	67.24
5	**36**	38.2	-2.2	4.84
6	**45**	38.2	6.8	46.24
7	**35**	38.2	-3.2	10.24
8	**45**	38.2	6.8	46.24
9	**35**	38.2	-3.2	10.24
10	**40**	38.2	1.8	3.24
Sum	**382**		0.0	209.6
Mean ($\overline{X}$)	**38.2**			

Now, we take these "squares" and sum them to get the Sum of Squares (SS) value. Here, the sum is 209.6. Next, we divide this sum by the number of scores minus 1.

Here, the result is $\frac{209.6}{10-1}$

= 23.29

This value is called the variance

The next step is to take the square root of the variance so as to obtain the standard deviation

Standard deviation = $\sqrt{23.2888}$ = 4.8258

Although this computation may seem so long and confusing, it is actually quite simple. To see this, consider the formula for the standard deviation:

$$\sqrt{\frac{\sum(X-\bar{X})^2}{\mathrm{n}-1}}$$

X = each score

$\bar{X}$= the mean or average

n = the number of values

$\sum$ means we sum across the values

In the top part of the ratio, the numerator, we see that each score has the mean subtracted from it, the difference is squared, and the squares are summed. In the bottom part, we take the number of scores minus 1. The ratio is the variance and the square root is the standard deviation. In English, we can describe the standard deviation as **the square root of the sum of the squared deviations from the mean divided by the number of observations minus one**

The standard deviation allows you to reach some conclusions about specific scores in your distribution. Assuming that the distribution of scores is normal or bell-shaped (or close to it!), the following conclusions can be reached:

Approximately 68% of the scores in the sample fall within one standard deviation of the mean

approximately 95% of the scores in the sample fall within two standard deviations of the mean

approximately 99% of the scores in the sample fall within three standard deviations of the mean

For instance, since the mean in our example is 38.2 and the standard deviation is 4.8258, we can from the above statement estimate that approximately 95% of the scores will fall in the range of 38.2-(2*4.8258) to 38.2+(2*4.8258)

or between 28.5484 and 47.8516.

This kind of information is a critical stepping stone to enabling you to compare the performance of an individual on one variable with their performance on another, even when the variables are measured on entirely different scales.

Tutor marked Assignment

2. What is central tendency? List the measure of central tendency
3. What is dispersion? What are the measures of dispersion:
4. What is the relationship between mean and variance?
5. What is the the differences between MAD and STD?
6. The egg production of 16 chickens to 30 weeks of age is given below.

67 57 56 57 58 56 54 64 53 54 54 55 57 68 60 58

Find the range, mean absolute deviation, variance, standard deviation, and coefficient of variation for this set of data.

4.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

Unit 6. Sampling, data collection and data processing techniques

6.1 Introduction

A research will first as a research question. The he will formulate hypothesis about the problem he want to solve. Then he decides on what kind of data should be gathered that can appropriately provide answer to his research question. In this unit you will learn how to source for and gathere your data.

6.2 Objective

At the end of this unit, you will know

Different methods of data collection

How to design questionnaire

Different type of sampling

6.3 Main Text

Where does data come from? How is it gathered? How do we ensure its accurate? Is the data reliable? Is it representative of the population from which it was drawn? We now explore some of these issues.

6.3.1 Methods of data collection

There are many methods used to collect or obtain data for statistical analysis. Three of the most popular methods are:

Direct Observation

Experiments, and

Surveys.

A survey solicits information from people; e.g. Village chicken production; pre-election polls; marketing surveys. The Response Rate (i.e. the proportion of all people selected who complete the survey) is a key survey parameter.

Surveys may be administered in a variety of ways, e.g. Personal Interview, Telephone Interview, and Self-Administered Questionnaire.

Questionnaire Design

Over the years, a lot of thought has been put into the science of the design of survey questions. Key design principles: 1. Keep the questionnaire as short as possible. 2. Ask short, simple, and clearly worded questions. 3. Start with demographic questions to help respondents get started comfortably. 4. Use dichotomous (yes | no) and multiple choice questions. 5. Use open-ended questions cautiously. 6. Avoid using leading-questions. 7. Pretest a questionnaire on a small number of people. 8. Think about the way you intend to use the collected data when preparing the questionnaire.

Sampling

You will recall that statistical inference permits us to draw conclusions about a population based on a sample. Sampling (i.e. selecting a sub-set of a whole population) is often done for reasons of cost (it's less expensive to sample 20 chickens out of 2000 especially when the carcass will not be useful after the experiment) and practicality (it is impossible to use the whole population for an experiment). In any case, the sampled population and the target population should be similar to one another.

Sampling methods

Sampling methods are classified as either probability or nonprobability.

In probability samples, each member of the population has a known non-zero probability of being selected. Probability methods include random sampling, systematic sampling, and stratified sampling.

In nonprobability sampling, members are selected from the population in some nonrandom manner. These include convenience sampling, judgment sampling, quota sampling, and snowball sampling. The advantage of probability sampling is that sampling error can be calculated. Sampling error is the degree to which a sample might differ from the population. When inferring to the population, results are reported plus or minus the sampling error.

In nonprobability sampling, the degree to which the sample differs from the population remains unknown.

Random sampling

Random sampling is the purest form of probability sampling. Each member of the population has an equal and known chance of being selected. When there are very large populations, it is often difficult or impossible to identify every member of the population, so the pool of available subjects becomes biased.

Systematic sampling

Systematic sampling is often used instead of random sampling. It is also called an Nth name selection technique. After the required sample size has been calculated, every Nth record is selected from a list of population members. As long as the list does not contain any hidden order, this sampling method is as good as the random sampling method. Its only advantage over the random sampling technique is simplicity.

Stratified sampling

Stratified sampling is commonly used probability method that is superior to random sampling because it reduces sampling error. A stratum is a subset of the population that share at least one common characteristic. Examples of stratums might be males and females, or breeds of chickens. The researcher first identifies the relevant stratums and their actual representation in the population. Random sampling is then used to select a sufficient number of subjects from each stratum. "Sufficient" refers to a sample size large enough for you to be reasonably confident that the stratum represents the population. Stratified sampling is often used when one or more of the stratums in the population have a low incidence relative to the other stratums.

Convenience sampling

Convenience sampling is used in exploratory research where the researcher is interested in getting an inexpensive approximation of the truth. As the name implies, the sample is selected because they are convenient. This nonprobability method is often used during preliminary research efforts to get a gross estimate of the results, without incurring the cost or time required to select a random sample.

Judgment sampling

Judgment sampling is a common nonprobability method. The researcher selects the sample based on judgment. This is usually an extension of convenience sampling. For example, a researcher may decide to draw the entire sample from one "representative" city, even though the population includes all cities. When using this method, the researcher must be confident that the chosen sample is truly representative of the entire population.

Quota sampling

Quota sampling is the nonprobability equivalent of stratified sampling. Like stratified sampling, the researcher first identifies the stratums and their proportions as they are represented in the population. Then convenience or judgment sampling is used to select the required number of subjects from each stratum. This differs from stratified sampling, where the stratums are filled by random sampling.

Snowball sampling

Snowball sampling is a special nonprobability method used when the desired sample characteristic is rare. It may be extremely difficult or cost prohibitive to locate respondents in these situations. Snowball sampling relies on referrals from initial subjects to generate additional subjects. While this technique can dramatically lower search costs, it comes at the expense of introducing bias because the technique itself reduces the likelihood that the sample will represent a good cross section from the population.

Sampling plan

A sampling plan is just a method or procedure for specifying how a sample will be taken from a population. We will focus our attention on these three methods: • Simple Random Sampling, • Stratified Random Sampling, and • Cluster Sampling.

Simple Random Sampling

A simple random sample is a sample selected in such a way that every possible sample of the same size is equally likely to be chosen. Drawing three names from a hat containing all the names of the students in the class is an example of a simple random sample: any group of three names is as equally likely as picking any other group of three names.

Stratified Random Sampling

A stratified random sample is obtained by separating the population into mutually exclusive sets, or strata, and then drawing simple random samples from each stratum.

Cluster sampling

is a simple random sample of groups or clusters of elements (vs. a simple random sample of individual objects). This method is useful when it is difficult or costly to develop a complete list of the population members or when the population elements are widely dispersed geographically. Cluster sampling may increase sampling error due to similarities among cluster members.

This is an important issue. Numerical techniques for determining sample sizes will be described later, but suffice it to say that the larger the sample size is, the more accurate we can expect the sample estimates to be.

Sampling error

Sampling error refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample. Another way to look at this is: the differences in results for different samples (of the same size) is due to sampling error: E.g. Two samples of size 10 of 1,000 broiler chickens weight at 8 weeks. If you happened to get the highest weights data points in your first sample and all the lowest weights in the second, this is a consequence of sampling error. Increasing the sample size will reduce this type of error.

Non-sampling error

Non-sampling error are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly.

There are three types of non-sampling errors:

1. Errors in data acquisition,

2. Nonresponse errors, and
3. Selection bias.

Increasing the sample size will not reduce this type of error.

Errors in Data Acquisition arise from the recording of incorrect responses, due to:

- incorrect measurements being taken because of faulty equipment,
- mistakes made during transcription from primary sources,
- inaccurate recording of data due to misinterpretation of terms, or
- inaccurate responses to questions concerning sensitive issues.

Non response Error refers to error (or bias) introduced when responses are not obtained from some members of the sample, i.e. the sample observations that are collected may not be representative of the target population. As mentioned earlier, the Response Rate (i.e. the proportion of all people selected who complete the survey) is a key survey parameter and helps in the understanding in the validity of the survey and sources of non response error.

Selection Bias occurs when the sampling plan is such that some members of the target population cannot possibly be selected for inclusion in the sample.

Data Processing Technique

Data processing methods are techniques used to process or sort different types of data

The commonly used data processing techniques are

Batch Processing

Online Processing

Real time processing

Multiprogramming

Multiprocessing and

Time sharing

Since data is being entered and processed immediately in real-time processing, the data can be accessed and corrected immediately by the user. Data that is processed in a batch must follow a structured protocol for correcting errors, which often takes more time. Real-time processing produces data that is more up-to-date than data processed in batches. It is also likely to produce more accurate data, since the input tools are readily available to users.

Batch processing can be more cost-effective, using fewer peripheral devices than real-time processing, though the cost savings is reduced as the price of peripheral devices decreases over time. Batch processing also allows a business to schedule when the computer is to be used, allowing for more efficient use of computer hardware and personnel time. Batches can be programmed to be processed at night, and are ready and waiting for workers the next morning.

5.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

Unit 7. Inference and hypothesis testing; Type I and type II errors

7.1 Introduction

7.2 Objective

You will be able to do the following after completing this unit

Define hypothesis testing

Define statistical inference

Formulate null and alternative hypothesis

Calculate the test statistics for hypothesis testing

Differentiate one tail from two tail test

Know the boundaries for rejection region for the hypothesis test

Differentiate between Type I error and type II error

7.3 Main Content

7.3.1 What is an Hypothesis?

An hypothesis is a statement or claim about a population. It is an assumption about a population parameter.

Example of hypotheses (plural) include the following

The weight of adult Bunaji bull is 300kg

The weight local chicken eggs in Nigeria is 30g

The average flock size of rural poultry farmer is 10 chickens

When you look at all of the statements made above any of them may be true and alternatively may be false.

The purpose of hypothesis testing is to help you determine which of the alternative hypothesis to accept.

Types of Statistical Inference

There are two common types of statistical inference

a) Confidence intervals if your goal is to estimate population parameter

b) test of significance, if your goal is to assess the evidence provided by a set of data about some claim concerning the population under examination

7.3.2 Formulating Null and Alternative Hypothesis

You will decide an make a statement you want to test. For example

There is no difference between the weight Bunaji Bull and Bunaji Cow at 1 year of age.

The alternative hypothesis is that the weight of Bunaji bull and cow differs at 1 year of age.

The alternative hypothesis is that things are different from each other, or different from a theoretical expectation.

The null hypothesis is written as H_0 while the alternative hypothesis is written as H_A or H_1.

The two examples of hypothesis above can be written as follows:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

You may also formulate your null hypothesis from your knowledge of the population. For example you know that average body weight of ram is 35kg. You can them make a statement about a West African Dwarf Ram. Your null hypothesis could be that the average weight of West African Dwarf Ram is 35kg (the known weight of rams from literature). The alternative hypothesis will then be that the average weight of West African Dwarf Ram is different from 35kg. This can be written using the mathematical notation:

H_0: $\mu = 35kg$

H_1: $\mu \neq 35kg$

The alternative hypothesis always states the mean of the population <, ≠ or > a specific value. In carrying out a test of H_0 versus H_1, H_0 will be rejected in favour of H_1 only if sample evidence strongly suggests that H_0 is false. If the sample does not provide such evidence, H_0 will not be rejected. So the 2 possible conclusions are reject H_0 and fail to reject H_0.

7.3.3 Test statistics for Hypothesis testing

The primary goal of a statistical test is to determine whether an observed data set is so different from what you would expect under the null hypothesis that you should reject the null hypothesis. Once you have formulated your hypothesis, you will carry out a test procedure. A test procedure is a method for making decision. This is based on a sample selected from the population under investigation. You will recall that statistic is any quantity whose value can be computed from a sample data. First select a particular statistic to serve as your decision maker. What is this decision? It is whether or not you should reject H_0. Whether you reject or you do not reject H_0 depends on the extent to which the value of this statistic computed from the sample is consistent with H_0. Generally speaking, if the computed value is very different from what would be expected when H_0 is true, rejection of H_0 is appropriate. However if the computed value is one that might reasonably have resulted when H_0 is true, then there is no strong reason to reject H_0 in favour of H_a.

Example:

Thirty (30) ShikaBrown layer chickens were randomly chosen from a population established at Poultry Research Programme farm at the National Animal Production Research Institute, Shika Zaria. The reported body weight of ShikaBrown reported over the year is 2500gm with a

standard deviation of 250gm at 40 weeks. Is this sample of 30 layers different from the population with 2500gm average weight at 40 weeks?

sno	body wt at 40 wks	Sno	body wt at 40 wks
1	2300	16	2500
2	2400	17	2500
3	2100	18	2200
4	2300	19	1950
5	2500	20	2200
6	2400	21	2350
7	1900	22	2150
8	1950	23	2200
9	1850	24	2800
10	2800	25	2500
11	1850	26	2500
12	2350	27	2150
13	2150	28	1850
14	2200	29	2100
15	2800	30	2000

From previous study, we know that the standard deviation is 250gm

The average body weight of this sample of 30 chickens is 2260g

Formulate the null hypothesis

H_0: $\mu = 2500$

H_a: $\mu \neq 2500$

The computed value of the sample mean of 2260gm is definitely different from the population mean of 2500gm. But is this difference big enough to reject the null hypothesis of no difference? The answer is highly dependent on the amount of variability in the body weight of the layer chickens at 40 weeks. If the population standard deviation σ is large then $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ will also b e large. In this case an $\bar{x}$ value rather far from μ would not be very unusual.

It is easier to take explicit account of variability in reaching a decision if a standardized statistic rather than $\bar{x}$ itself is used as a decision maker. The sample size is less than 30 so you will use a t statistic. The standard deviation of $\bar{x}$ is $\sigma_{\bar{x}} = \frac{250}{\sqrt{30}} = 45.64$

When H_0 is true, $\mu_{\bar{x}}$=2500.

Standardizing $\bar{x}$ assuming H0 is true yields the standardized statistic

$$z = \frac{\bar{x}-\text{hypothesized value}}{\text{standard deviation of } \bar{x}} = \frac{\bar{x}-2500}{45.64} = \frac{2260-2500}{45.64} = -5.25$$

You will now determine the critical z score which correspond to α=0.05. Because this is a two-tail test, this area needs to be evenly divided between both tails with each tail receiving $\alpha/2 = 0.05/2 = 0.025$.

You need to find the critical z-score that corresponds to the area 0.950+0.025 = 0.975. The area 0.950 is derived from 1- α. Using the z table in the appendix you look for the closest value to 0.9750 in the body of the table we can find this value looking across column 1.9 and down row 0.06 to arrive at the z-score of 1.96 for the right tail and -1.96 for the left tail. You can use either the scale of the original variable or the standardized normal scale to make a decision to reject or not to reject the null hypothesis.

Using the scale of the original variable

This section will determine the rejection region using the scale of the original variable, which in this case is the body weight of layer at 40 weeks. To calculate the upper and lower limits of the rejection region, we use the following equations. (Remember that we use the z-scores from standard normal distribution when n ≥ 30 and σ is known.

Limits of rejection region = $\mu_{H_0} + z_c\sigma_{\bar{x}}$

Where μ_{H_0}= population mean assumed by the null hypothesis

For our example

Upper limit = $\mu_{H_0} + z_c\sigma_{\bar{x}}$ = 2500 + 1.96(45.64)=2589.45

Lower Limit = $\mu_{H_0} + z_c\sigma_{\bar{x}}$ = 2500 + 1.96(45.64)=2410.55

Because our sample mean is 2260gm, this falls out the "do not reject region".

Your conclusion is that the difference between 2500 and 2260 gm is large enough to be due to chance variation. Therefore we reject the null hypothesis and accept the alternative.

Using the standardized Normal Scale

You can arrive at the same conclusion by setting up the boundaries for the rejection region using the standardized normal scale. We do this by calculating the z-score that corresponds to the sample mean as follows

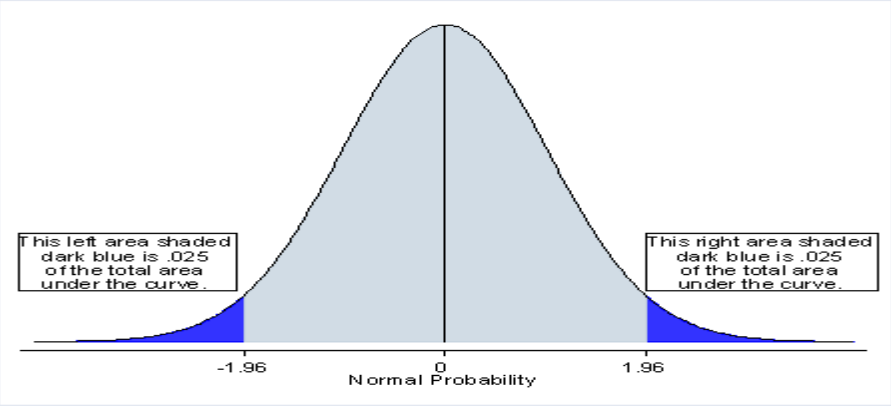
$z = \frac{\bar{x} - \mu_{H_0}}{\sigma_{\bar{x}}} = \frac{2260 - 2500}{45.64}$ = -5.26

Because the calculated value of -5.26 is outside the "Do not reject H_0" region, the conclusions based on this second technique is consistent with the first technique.

7.3.4 One tail versus two-tailed test

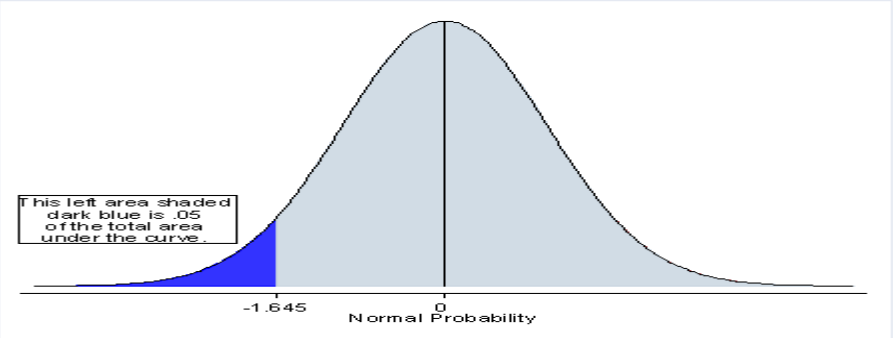
First let's start with the meaning of a two-tailed test. If you are using a significance level of 0.05, a two-tailed test allots half of your alpha to testing the statistical significance in one direction and half of your alpha to testing statistical significance in the other direction. This means that .025 is in each tail of the distribution of your test statistic. When using a two-tailed test, regardless of the direction of the relationship you hypothesize, you are testing for the possibility of the relationship in both directions. For example, we may wish to compare the mean of a sample to a given value *x* using a t-test. Our null hypothesis is that the mean is equal to *x*. A two-tailed test

will test both if the mean is significantly greater than *x* and if the mean significantly less than *x*. The mean is considered significantly different from *x* if the test statistic is in the top 2.5% or bottom 2.5% of its probability distribution, resulting in a p-value less than 0.05.



What is a one-tailed test?

Next, let's discuss the meaning of a one-tailed test. If you are using a significance level of .05, a one-tailed test allots all of your alpha to testing the statistical significance in the one direction of interest. This means that .05 is in one tail of the distribution of your test statistic. When using a one-tailed test, you are testing for the possibility of the relationship in one direction and completely disregarding the possibility of a relationship in the other direction. Let's return to our example comparing the mean of a sample to a given value *x* using a t-test. Our null hypothesis is that the mean is equal to *x*. A one-tailed test will test either if the mean is significantly greater than *x* or if the mean is significantly less than *x*, but not both. Then, depending on the chosen tail, the mean is significantly greater than or less than *x* if the test statistic is in the top 5% of its probability distribution or bottom 5% of its probability distribution, resulting in a p-value less than 0.05. The one-tailed test provides more power to detect an effect in one direction by not testing the effect in the other direction. A discussion of when this is an appropriate option follows.



Example of one tail test

The weight local chicken eggs is know to be less than that of exotic chickens. A sample of 30 eggs was obtain from a population of local chickens. The egg size of local chickens in Nigeria obtained from literature is 40 g with standard deviation of 10g. The average of the sampled 30 eggs is 35g. Is the sampled average egg weight less than the population egg weight?

H_0: $\mu < 40$

H_a: $\mu > 40$

Where μ = the mean egg weight in gram

The standard deviation of the population mean is 10g

Sample size n is 30 and $\alpha = 0.05$

The standard error of the mean $\sigma_{\bar{x}} == \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{30}} = 1.8257$

The sample mean from 30 eggs is 35g. What is our conclusion about our estimate of the population mean, μ?

The next step is for you to determine the critical z-score which correspond to $\alpha = 0.05$. Because this is a one tail test, this entire area needs to be in one rejection region on the right side of the distribution. According to figure ---- you need to find the z-score that corresponds to the area 0.95 or 1- α.

Using the z table in the appendix you will look for the closest value to 0.9500 in the body of the table, which results in a critical z-score of 1.65.

To calculate the limit for this rejection region using the scale of the original variable we use

Upper limit = $\mu_{H_0} + z_c\sigma_{\bar{x}}$ = 40 + 1.65(1.8257)=43.01gm

Since our mean is 35g and fall within do not reject region. Therefore we did not reject the null hypothesis that the local egg weight is significantly less than 40g.

Using standardized scale,

$z = \frac{\bar{x}-\mu_{H_0}}{\sigma_{\bar{x}}} = \frac{35-40}{1.8257}$ = -2.74

As you can see the calculated z-score is within the "Do not reject Region H_0" is consistent with your earlier finding.

7.3.5 Error type

Type I error

When the null hypothesis is true and you reject it, you make a type I error. The probability of making a type I error is α, which is the level of significance you set for your hypothesis test. An α of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. To lower this risk, you must use a lower value for α. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists.

Type II error

When the null hypothesis is false and you fail to reject it, you make a type II error. The probability of making a type II error is β, which depends on the power of the test. You can

decrease your risk of committing a type II error by ensuring your test has enough power. You can do this by ensuring your sample size is large enough to detect a practical difference when one truly exists.

Tutor marked assignment

1. What is inference statistics? What is hypothesis testing?
    a. What is an hypothesis?
    b. Hypotheses are normally formulated in pairs.
        i. What is the first hypothesis called?
        ii. What is the commonly name given to the second hypothesis?
2. In Hypothesis testing certain quantity must be calculated
    a. What quantity z or t do you need to calculate for large sample and how large should the sample be.
    b. What are the 2 major conditions to be satisfied before you can use z statistics?
    c. What is the formula for calculating z statistics?
3. Formulate hypothesis statement for the following claim: "The average Bunaji cow produces 6kg of milk daily." A sample of 40 Bunaji cows produced an average 8 kg of milk per day. Assume the population standard deviation is 2.5kg.Using $\alpha = 0.05$, test your hypothesis. What is your conclusion?
4. Formulate hypothesis statement for the following claim: "The average age at first egg of Nigerian breed of chicken called ShikaBrown is less than 150 days." A sample of 50 chickens had an average age at first egg of 140 days. Assume the population standard deviation is 22 days. Using $\alpha = 0.05$, test your hypothesis. What is your conclusion?
5. In Hypothesis testing, you can make two major types of error. List the types of error and describe them adequately.

6.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

Unit 8. Analysis of Variance
8.1 Introduction

Earlier in the early module, you learnt about two sample z- and t-tests for the difference between two conditions of an independent variable. We often need a tool for comparing more than two sample means.

In this unit, you will learn about a new parametric statistical procedure to analyze experiments with two or more conditions of a single independent variable. Then, you will learn about application of this technique to more than one independent variable.

ANalysis Of Variance = ANOVA is a very popular inferential statistical procedure It can be applied to many different experimental designs Independent or related samples. An independent variable with any number of conditions, or levels may be involved

Recall from our earlier lecture on experimental design

A one-way: ANOVA is performed when there is only one independent variable

One thing this means is we will be looking for a significant difference in mean, but we'll do it by looking at a ratio of variances

Assumptions of ANOVA

The dependent variable is quantitative The data was derived from a random sample The population represented in each condition is distributed according to a normal distribution

The variances of all the populations are homogenous

It is not required that you have the same number of samples in each group, but ANOVA will be more robust to violations of some of its other assumptions if this is true.

8.2 Objective

At the end of this unit, you will

1. be able to understand the need to analyse data from two samples
2. understand the underlying models to ANOVA
3. know the steps in performing a one-way anova
4. be able to interpret the result of one-way ANOVA

8.3 Main Text

8.3.1 Analysis of differences between more than two means

Many investigations comprises of comparison of more than two population or treatment means The characteristic that distinguishes the populations or treatments from one another is called the **factor** under investigation. In your example, the factor is breed. A single factor analysis of variance (ANOVA) problem involves a comparison of k population or treatment means μ_1, μ_2, μ_3,…. μ_k. The objective is to test :

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ (no difference in the true mean body weights at marketing)

Against

H_a: at least two of the means are different

The analysis is based on k independently selected random samples one from each population or for each treatment. That is, in case of populations, the sample from any particular population is selected independently of that from any other population. When the experimental units (subjects or objects) that receive any particular treatment are chosen independently of the units that receive any other treatment. This is known as completely randomised design, a comparison of treatment based on independently selected experimental units.

8.3.2 Underlying models to ANOVA

The normally used normal linear models for a completely randomized experiment are

$y_{i,j} = \mu_i + \mathcal{E}_{i,j}$ (the mean model)

or

$y_{i,j} = \mu + T_j + \mathcal{E}_{i,j}$ (the effect model)

Where

i = 1, …., i is an index over experimental units

j = 1, …., j is an index over treatment groups

I_j = is the number of experimental units in the j^{th} treatment group

$I = \sum_j I_j$ is the total number of experimental units

$y_{i,j}$ are observations

μ_j *i*s the mean of the observations for j^{th} treatment group

μ_j *i*s the grand mean of the observations

T_j is the j^{th} treatment effect, a deviation from the grand mean

$=\sum T_j = 0$

$\mu_j = \mu + T_j$

$\mathcal{E} \sim N(0,\sigma^2)$ $\mathcal{E}_{ij}$ are normally distributed zero-mean random errors.

You will use one-way ANOVA to study the effect of a single factor at more than 2 levels.

You will determine if the different levels of a single factor affects a measured observation. To do this , you must formulate appropriate hypotheses as follow

H_0: $\mu_i = \mu$ all i = 1, 2, … k

H_a: $\mu_i \neq \mu$ some i = 1, 2, … k

Where μ is the population mean for level i

Assumption of ANOVA

1. The observations are obtained independently and randomly from populations defined by the factor levels
2. The population at each factor levels is approximately normally distributed.
3. These normal populations have a common variance σ^2.

8.3.4 Steps in performing a one-way anova

1. Decide whether you are going to do a Model I or Model II ANOVA.
2. If you are going to do a Model I ANOVA, decide whether you will do planned comparisons of means or unplanned comparisons of means. A planned comparison is where you compare the means of certain subsets of the groups that you have chosen in advance.
3. If you are going to do planned comparisons, decide which comparisons you will do. If you are going to do unplanned comparisons, decide which technique you will use.
4. Collect your data.
5. Make sure the data do not violate the assumptions of the anova (normality and homoscedasticity) too severely. If the data do not fit the assumptions well enough, try to find a data transformation that makes them fit. If this doesn't work, do a Welch's ANOVA or a Kruskal–Wallis test instead of a one-way ANOVA.
6. If the data do fit the assumptions of an ANOVA, test the heterogeneity of the means.
7. If you are doing a Model I ANOVA, do your planned or unplanned comparisons among means.

8. If the means are significantly heterogeneous, and you are doing a Model II anova, estimate the variance components (the proportion of variation that is among groups and the proportion that is within groups).

A study is conducted to determine the differences among dairy cows in milk yield that is due to different herds. A random sample of cows from a random sample of herds chosen among all herds is measured to determine if differences among means are large enough to conclude that herds are generally different. This example demonstrates a random effects model because the herds measured are a random sample of all possible herds.

In applying a completely randomized design or when groups indicate a natural way of classification, the objectives may be:

1. Estimating the means

2. Testing the difference between groups

Although computer programs that do ANOVA calculations are now very common, it is very good that you know how to calculate the various entries in an ANOVA table.

The goal of ANOVA is to produce two variances (treatments and error) and their ratio.

8.3.5 Steps in ANOVA Calculation

Step 1 Compute Correction factor (CF)

$$CF=\frac{\left(\sum_{i=1}^{3}\sum_{j=1}^{5}y_{ij}\right)^2}{N_{total}}=\frac{(Total\ of\ all\ observation)^2}{N_{total}}$$

Step 2. Compute total SS

The total sum of square (SS) =SS(Total) = sum of square of all observation – CF

$$SS(Total)=\sum_{i=1}^{3}\sum_{j=1}^{5}y_{ij}^{2}-CF$$

Step 3. Compute SST, the treatment sum of squares

First you will compute the total (sum) for each treatment

Then SST=$\sum_{i=1}^{3} \frac{T_i^2}{n_i}$ - CF

Step 4. Compute SSE the error sum of square

SSE = SS(Total) – SST

Step 5 Compute MST, MSE and their ratio, F

MST is the mean square of treatments, MSE is the mean square of error (MSE is frequently denoted as $\hat{\sigma}_e^2$)

MST =$\frac{SST}{k-1}$

MSE =$\frac{SSE}{n-k}$

Where n is the total number of observation and k is the number of treatments finally compute F as

F =$\frac{MST}{MSE}$

Now use the steps listed above to produce an ANOVA table for the following data.

The data below resulted from measuring the difference in daily weight gain resulting in feeding three group of chickens three different diets with different levels of protein. Five chickens were randomly allocated to each of the 3 group. In short, the design of the experiment is such that you have an experiment in which each of the three treatments was replicated 5 times.

Level1	Level 2	Level 3
6.9	8.3	8.0
5.4	6.8	10.5
5.8	7.8	8.1
4.6	9.2	6.9
4.0	6.5	9.3

Step 1. Compute the correction factor

$$CF=\frac{\left(\sum_{i=1}^{3}\sum_{j=1}^{5}y_{ij}\right)^2}{N_{total}}=\frac{(\text{Total of all observation})^2}{\text{N}_{\text{total}}}=\frac{108.1^2}{15}=779.041$$

$$SS(Total)=\sum_{i=1}^{3}\sum_{j=1}^{5}y_{ij}^2-CF$$

$= (6.9)^2 + (5.4)^2 + \ldots + (6.9)^2 + (9.3)^2 - CF$

$= 829.390 - 779.041 = 45.439$

Step 3. Compute SST, the treatment sum of squares

First you will compute the total (sum) for each treatment

$T_1 = 6.9 + 5.4 + 5.8 + 4.6 + 4.0 = 26.7$

$T_2 = 8.3 + 6.8 + 7.8 + 9.2 + 6.5 = 38.6$

$T_3 = 8.0 + 10.5 + 8.1 + 6.9 + 9.3 = 42.8$

Then $SST=\sum_{i=1}^{3}\frac{T_i^2}{n_i} - CF = \frac{(26.7)^2}{5}+\frac{(38.6)^2}{5}+\frac{(42.8)^2}{5} - 779.041 = 27.897$

Step 4. Compute SSE the error sum of square

SSE = SS(Total) – SST = 45.349 – 27.897 = 17.45

Step 5 Compute MST, MSE and their ratio, F

MST is the mean square of treatments, MSE is the mean square of error (MSE is frequently denoted as $\hat{\sigma}_e^2$)

$$MST=\frac{SST}{k-1}=\frac{27.897}{2}=13.949$$

$$MSE=\frac{SSE}{N-k}=\frac{17.452}{12}=1.454$$

Where N is the total number of observations and k is the number of treatments. Finally, compute F as show bellow

$F = \frac{MST}{MSE} = \frac{13.949}{1.454} = 9.59$

You will now go ahead to assemble this number in an ANOVA Table.

8.3.6 The ANOVA Table and tests of hypotheses about means

You have computed two sums of squares SST and SSE (for a one-way ANOVA). From these sums of squares you have also computed two mean squares one for treatment and the other for error. They are denoted as MST and MSE respectively. You will now display the values in a table referred to as ANOVA Table. The table will also display the statistics used to test hypotheses about the population means.

When the null hypothesis of equal means is true, the two mean squares estimate the same quantity (error variance) and should be of approximately equal magnitude. In order words their ratio should be close to 1. If the null hypothesis is false, MST should be larger than MSE

The mean squares are formed by dividing the sum of squares by the associated degrees of freedom.

Let $N = \Sigma n_i$ and k is the number of treatment

Then the degree of freedom for treatment are

DFT = k – 1

And degree of freedom for error (DFE) = N-k

The corresponding mean squares are:

MST = SST/DFT

MSE = SSE/DFE

The test statistic used in testing the equality of treatment means is

F = MST/MSE

The critical value is the tabular value of the F distribution based on the chosen α level and the degree of freedom DFT and DFE.

The calculations are displayed in an ANOVA Table as follows:

Source	SS	DF	MS	F
Treatment	SST	k-1	SST/(k-1)	MST/MSE
Error	SSE	N-k	SSE/(N-k)	
Total (corrected)	SS	N-1		

The word source stands for source of variation. You may also find that treatment is replace by between and Error by within in some textbooks.

The ANOVA table for your example will look like the following table.

Source	SS	DF	MS	F
Treatment	27.897	2	13.949	9.59
Error	17.452	12	1.454	
Total (corrected)	45.349	14		
Correction Factor	779.041	1		

8.3.7 Interpretation of ANOVA Table

The test statistic is the F value of 9.59. Using an α of 0.05 we have $F_{0.05;\,2,\,12} = 3.89$ (You can get F tables from any standard Statistics textbook). Since the test statistic is much larger than the critical value, we reject the null hypothesis of equal population means and conclude that there is a (statistically) significant difference among the population means. The p-value for 9.59 is 0.00325. So the test statistics is significant at that level.

At this point what we know is that the means are different. However we don't know whether only one of the mean is different from the other two or all of the three means are different from one another.

8.3.8 Making multiple comparison

Once the null hypothesis is rejected, the next stage is to determine which of the treatment means is different from each other. You may also ask the question such "Does the average of treatment 1 and 2 differ significantly from the average of treatment 3 and 4?"

To answer such question you need to proceed to multiple comparison procedures. There are many of them.

You can use any of the following procedure to determine differences in the means. This is also known as Separation of Means. The popular ones are:

Tukey's Method to test all possible pairwise differences of means to determine if at least one difference is significantly different from 0.

Sheffe's Method to test all possible contrasts at the same time, to see if at least one is significantly different from 0.

Bonferroni Method to test, or put simultaneous confidence intervals around, a pre-selected group of contrast.

However you will learn later how to use some computer programs to separate your different means. Meanwhile you may use the simplest method of Bonferroni described below to separate your means as it is the simplest of the methods.

One way to determine specific difference is to perform paired analyses of the group, two at a time. For example, compare the mean for group A vs the mean for group B, then A vs C then A vs D and so on.

If you do these pairwise comparisons, you should modify the resulting p-value for each t-test, since performing multiple t-tests increases the probability of finding an incorrect significance. To correct for this problem you should multiply the p-values for each of the pair-wise comparisons

by the number of comparisons. This is called a Bonferonni adjustment. For example, in this case your comparisons are

A vs B, A vs C, A vs D, B vs C, B vs D, and C vs D -- 6 pairwise comparisons in all. Thus, you'd correct each t-test p-value by multiplying it by 6.

For example, a t-test comparison of Mean A vs Mean C (61.025 vs 89.067) yields an unadjusted two-tail p-value p=0.0006. The adjusted p-value (the one you should report) would be 0.0006*6 = 0.0036

8.3.9 Another example of one-way anova

The following table contains information on four different feeds and weight gain of animals after they had been fed one of the feeds for a period of time. You want to know if any feed is better for producing weight gain. I will use a computer programme to solve this.

Feed type			
A	B	C	D
60.8	78.7	92.6	86.9
67.0	77.7	84.1	82.2
54.6	76.3	90.5	83.7
61.7	79.8	100.2	90.3

8.3.10 The two-way ANOVA

This is probably the most popular layout in the design of experiments. This is also refer to as factorial experiment

So can you define a factorial experiment?

An experiment that utilizes every combination of factor levels as treatments is called a factorial experiment.

8.3.11 Model for two-factorial experiment

In a factorial experiment with factor A at a levels and factor B at b levels, the model for the general layout can be written as

$Y_{ij} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$

For $i = 1,2, \ldots, a$

$j = 1,2, \ldots, b$

$k = 1,2, \ldots, r$

Where μ is the overall mean response, τ_i is the effect due to the i^{th} level of factor A, β_j is the effect due to the j^{th} level of factor B, and γ_{ij} is the effect due to any interaction between the i^{th} level of factor A and j^{th} level of factor B.

At this point, consider the levels of factor A and of factor B chosen for the experiment to be the only levels of interest to the experimenter such as predetermined levels for for feeding broilers or breed to which the chicken belong. The factors A and B are said to be fixed factors and the model is a fixed-effects model.

When an a×b factorial experiment is conducted with an equal number of observations per treatment combination, the total (corrected) sum of squares is partitioned as:

SS(total) = SS(A)+SS(B)+SS(AB)+SSE,

where AB represents the interaction between A and B.

For reference, the formulas for the sums of squares are:

$SS(A) = rb \sum_{i=1}^{a} (\bar{y}_{i..} - \bar{y}_{...})^2$

$SS(B) = ra \sum_{j=1}^{b} (\bar{y}_{.j.} - \bar{y}_{...})^2$

$SS(AB) = \sum_{i=1}^{b} \sum_{j=1}^{a} (\bar{y}_{ij.} - \bar{y}_{i..} + \bar{y}_{.j.} + \bar{y}_{...})^2$

$SSE = \sum_{k=1}^{r} \sum_{i=1}^{b} \sum_{j=1}^{a} (y_{ijk} - \bar{y}_{ij.})^2$

SS(Total) $= \sum_{k=1}^{r} \quad \sum_{i=1}^{b} \quad \sum_{j=1}^{a} (y_{ijk} - \bar{y}_{...})^2$

The resulting ANOVA table for an a x b factorial experiment

Source	SS	DF	MS
Factor A	SSA	a-1	SSA/(a-1)
Factor B	SSB	b-1	SSB/(b-1)
Interaction	SS(AB)	(a-1)(b-1)	SS(AB)/(a-1)(b-1)
Error	SSE	(N-ab)	SSE/(N-ab)
Total (Corrected	SS(Total)	(N-1)	

The various hypotheses that can be tested using this ANOVA table concern whether the different levels of Factor A, or Factor B, really make a difference in the response, and whether the AB interaction is significant.

For the two-way ANOVA, the possible **null hypotheses are**:

There is no difference in the means of factor A

There is no difference in means of factor B

There is no interaction between factors A and B

The **alternative hypothesis** for cases 1 and 2 is: the means are not equal.

The alternative hypothesis for case 3 is: there is an interaction between A and B.

Example:

An experiment was conducted to determine the effect of adding two vitamins (I and II) in feed on average daily gain of pigs. Two levels of vitamin I (0 and 4 mg) and two levels of vitamin II (0 and 5 mg) were used. The total sample size was 20 pigs, on which the four combinations of vitamin I and vitamin II were randomly assigned. The following daily gains were measured:

Vitamin I	0 mg		4mg	
Vitamin II	0 mg	5 mg	0 mg	5 mg
	0.585	0.567	0.473	0.684
	0.536	0.545	0.450	0.702
	0.458	0.589	0.869	0.900
	0.486	0.536	0.473	0.698
	0.536	0.549	0.464	0.693
Sum	2.601	2.786	2.729	3.677
Average	0.520	0.557	0.549	0.735

Total sum:

$\Sigma i\ \Sigma j\ \Sigma k\ yijk$ = (0.585 + ... + 0.693) = 11.793

Correction factor

$CF = \frac{(\Sigma i\ \Sigma j\ \Sigma k\ yijk)^2}{abn} = \frac{(11.793)^2}{20} = 6.953742$

3) Total sum of squares

$\Sigma_i\Sigma_j\Sigma_k(y_{ijk})^2 - CF = 0.585^2 + 0.536^2 + \ldots . 0.698^2 + 0.693^2 - 6.953742$

= 0.32169455

3) Sum of squares for Vitamin 1

$SS_{vit1} = \sum_i \frac{(\Sigma j\ \Sigma k\ yijk)^2}{nb} - CF = \frac{(2.601+2.786)^2}{10} + \frac{(2.729+3.677)^2}{10} - 6.953742$

= 0.05191805

3) Sum of squares for Vitamin 2

$SS_{vit2} = = \sum_j \frac{(\Sigma i\ \Sigma k\ yijk)^2}{na} - CF = \frac{(2.601+ 2.729)^2}{10} + \frac{(3.677+2.786\ )^2}{10} - 6.953742$

= 0.06418445

6) Sum of squares for interaction:

$SS_{vit1\ X\ Vit2} \sum_i \sum_j \frac{(\Sigma k\ yijk)^2}{n} - SSA - SSB - CF$

$= \frac{(2.601)^2}{5} + \frac{(\ 2.729)^2}{5} + \frac{(3.677)^2}{5} + \frac{(2.786\ )^2}{5}$ - 0.05191805 - 0.06418445 – 6.953742

= 0 02910845

7) Residual sum of squares

SSRES = SSTOT – SSVit I – SSVit II - SSVit I x Vit II

= 0.32169455 – 0.05191805 – 0.06418445 – 0.02910845

= 0.17648360

The *ANOVA* table is:

Source	*SS*	*df*	*MS*	*F*
Vitamin I	0.05191805	1	0.05191805	4.71
Vitamin II	0.06418445	1	0.06418445	5.82
Vit I x Vit II	0.02910845	1	0.02910845	2.64
Residual	0.17648360	16	0.01103023	
Total	0.32169455	19		

The critical value for $\alpha = 0.05$ is $F_{0.05,1,16} = 4.49$. The computed *F* value for the interaction is 2.64. In this case the calculated *F* value is less than the critical value.

Tutor marked assignment

There three sources of protein for feeding swine in Zaria. You were asked suggest if performance of swine when fed the different diet formulated using different sources of protein is the same. Fifteen growing pigs are available for this experiment. Five pigs were randomly allocated to each of the three diets and the result is tabulated as follows

Pigs	Diet		
	I	II	III
1	25	10	15
2	15	15	20
3	20	15	25
4	15	12	20
5	20	10	20

1. Is growth performance of pigs significantly related to source of protein?

2. What is the difference in the mean growth performance between each of the protein source

3. Which group means are significantly different from the mean for Diet I?

2. Four breeds were available for farmer for egg production. Determine which of this breeds is the best for egg production given the following experimental results

lines	Observation egg production					
1	79	62	66	86	89	101
2	57	75	98	61	84	96
3	68	75	50	74	53	61
4	64	71	79	45	50	40

a) Write out your null hypothesis for the above situation and the alternative.

b) What is your test statistics (One way anova)

c) Your degree of freedom k=4-1 and N-k = 24-4 =20 at $\alpha = 0.05$

d) Compute the total sum of square

e) Compute the treatment sum of square

f) Compute the Error sum of square

g) Compute the mean square for treatment and error and the F ratio

h) Compare the computed F ratio with the critical F ratio at df 3, 20 and $\alpha = 0.05$

8.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

Unit 9. Correlation and regression analysis

9.1 Introduction

Is the amount of milk produce by a cow related to the weight of her calf at weaning? Is the level of feeding of broiler chicken related to the weight of the broiler chicken at 8 weeks when it should be slaughter for marketing? At the end of this unit, you will be able to quantify your answer to questions of this type based on the data you might have gathered.

Correlation and regression are other areas of inferential statistics which involve determining whether a relationship between two or more numerical or quantitative variables exists. This is when two characteristics are studied simultaneously on each member of a population in order to examine whether they are related. For instance, a researcher may be interested in finding out the relationship between weight and age of broiler chickens or Lactation length of cows and weaning weight of the calves

Therefore, correlation and regression analyses are used to measure association between two variables of a bivariate data.

9.2 Objectives

At the end of this unit, you will be able to:

calculate the strength and direction of a relationship between two variables by collecting measurements and using suitable statistical analysis

evaluate and interpret the product moment correlation coefficient and Spearman's correlation coefficient

find the equations of regression lines and use them where appropriate

Define correlation and regression and bivariate distribution.

Explain the types of correlation and regression

State possible relationships between variables

Draw a scatter diagram for a bivariate data

Compute correlation and regression.

9.3 Main Content

9.3.1 Correlation

You often wonder what is the relationship between the height of an egg and its weight . Now go to take 10 eggs. Measure the height of each egg and its weight. Does there appear to be connection between the height and weight of the eggs?

Correlation is a statistical measure that indicates the extent to which two or more variables drawn from the same population fluctuate together. Correlation coefficient calculated fron a sample data measures the strength and direction of a linear relationship between two variables A positive **correlation** indicates the extent to which those variables increase or decrease; a negative **correlation** indicates the extent to which one variable increases as the other decreases. The symbol of correlation coefficient calculated from a sample data is r while the symbol for population correlation coefficient is ρ (rho). The relationship between two variables are not perfect.

9.3.1 Techniques in determining correlation

Pearson's Product-Moment Correlation is the most commonly used technique. Correlation is only appropriate for quantifiable data in which numbers are meaningful. You cannot use it for purely categorical data such as gender and brand.

9.3. Correlation coefficient

The formula for calculating r is

$$r = \frac{N\sum XY - (\sum X \sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

Where n is the number of data pairs (X, Y)

Or

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

9.3.2 Computational Procedure for simple linear correlation of correlation

Compute the mean x and y, the corrected sum of squares, $\sum x^2$ and $\sum y^2$ and the corrected sum of cross products $\sum xy$, of the variables x and y.

2. Compute the r value.

3. Compare the absolute value of the computed r value to the tabular r values with n-2 degrees of freedom at 5% and 1% levels of significance.

4. If the computed r value is greater than the tabular r value at 5% level but smaller than the tabular r value at the 1% level, then the simple linear correlation is significant at the 5% level of significance.

The range of 'r' value

The range of values for correlation coefficient is from -1 to +1 that is if there is 1.

Strong positive linear relationship between the variables, the value or r will be close to +1.

2. Strong negative linear relationship between the variables, the value of r will be close to -1.

3. No linear relationship between the variables or only a weak relationship, the value of r will be close to 0.

A key thing to remember when working with correlations is never to assume a correlation means that a change in one variable causes a change in another. Sales of personal computers and athletic shoes have both risen strongly in the last several years and there is a high correlation

between them, but you cannot assume that buying computers causes people to buy athletic shoes (or vice versa).

The second thing is that the Pearson correlation technique works best with linear relationships: as one variable gets larger, the other gets larger (or smaller) in direct proportion. It does not work well with curvilinear relationships (in which the relationship does not follow a straight line). An example of a **curvilinear relationship** is age and health care. They are related, but the relationship doesn't follow a straight line.

Suppose that we took 7 chickens and measured their body weight and their length from beak to tail. We obtained the following results and want to know if there is any relationship between the measured variables. [To keep the calculations simple, we will use small numbers]

Chicken	Units of weight (x)	Units of length (y)
1	1	2
2	4	5
3	3	8
4	4	12
5	8	14
6	9	19
7	8	22

Procedure for computation

Plot the results on graph paper. This is the essential first step, because only then can we see what the relationship might be - is it linear, logarithmic, sigmoid, etc?

In our case the relationship seems to be linear, so we will continue on that assumption. If it does not seem to be linear we might need to transform the data.

(2) Set out a table as follows and calculate S x, S y, S x2, S y2, S xy, and (mean of y).

	Weight (x)	Length (y)	x2	y2	xy
Mouse 1	1	2	1	4	2
Mouse 2	4	5	16	25	20
Mouse 3	3	8	9	64	24
Mouse 4	4	12	16	144	48
Mouse 5	8	14	64	196	112
Mouse 6	9	19	81	361	152
Mouse 7	8	22	64	484	176
Total	$S_x = 37$	$S_y = 82$	$S_x^2 = 251$	$S_y^2 = 1278$	$S_{xy} = 553$
Mean	= 5.286	= 11.714			

$$\sum d_x{}^2 = \sum x^2 - \frac{(\sum x)^2}{n} = 1^2 + 4^2 + \ldots + 8^2 - 37^2/7 = 55.429$$

$$\sum d_y{}^2 = \sum y^2 - \frac{(\sum y)^2}{n} = 2^2 + 5^2 + \ldots + 22^2 - 82^2/7 = 317.429$$

$$\sum d_x\, d_y = \sum xy - \frac{\sum x \sum y}{n} = 119.571$$

6) Calculate r (correlation coefficient)

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x{}^2 \sum d_y{}^2}} = \frac{119.571}{\sqrt{55.429\ X\ 317.429}} = 0.9014$$

r= 0.9014

(7) Look up r in a table of correlation coefficients (ignoring + or - sign). The number of degrees of freedom is two less than the number of points on the graph (5 df in our example because we have 7 points). If our calculated r value exceeds the tabulated value at p = 0.05 then the correlation is significant. Our calculated value (0.9014) does exceed the tabulated value (0.754). It also exceeds the tabulated value for p = 0.01 but not for p = 0.001. If the null hypothesis were true (that there is no relationship between length and weight) we would have obtained a

correlation coefficient as high as this in less than 1 in 100 times. So we can be confident that weight and length are positively correlated in our sample of mice.

Important notes:

9. If the calculated r value is positive (as in this case) then the slope will rise from left to right on the graph. As weight increases, so does the length. If the calculated value of r is negative the slope will fall from left to right. This would indicate that length decreases as weight increases.
10. The r value will always lie between -1 and +1. If you have an r value outside of this range you have made an error in the calculations.
11. Remember that a correlation does not necessarily demonstrate a causal relationship. A significant correlation only shows that two factors vary in a related way (positively or negatively). This is obvious in our example because there is no logical reason to think that weight influences the length of the animal (both factors are influenced by age or growth stage). But it can be easy to fall into the "causality trap" when looking at other types of correlation.

What does the correlation coefficient mean?

The part above the line in this equation is a measure of the degree to which x and y vary together (using the deviations d of each from the mean). The part below the line is a measure of the degree to which x and y vary separately.

9.3.3 Regression

Correlation and **regression** analysis are related in the sense that both deal with relationships among variables. We can use the technique of correlation to test the statistical significance of the

association. In other cases we use regression analysis to describe the relationship precisely by means of an equation that has predictive value .

You can fit a line to the data we have just analysed - producing an equation that shows the relationship, so that you might predict the body weight of chicken by measuring their length, or vice-versa. The method for this is called linear regression.

However, this is not strictly valid because linear regression is based on a number of assumptions. In particular, one of the variables must be "fixed" experimentally and/or precisely measureable. So, the simple linear regression methods can be used only when we define some experimental variable (temperature, pH, dosage, etc.) and test the response of another variable to it.

The variable that we fix (or choose deliberately) is termed the independent variable. It is always plotted on the X axis. The other variable is termed the dependent variable and is plotted on the Y axis.

Assumptions of the Regression Model

the relation between x and y is given by

$y = a + b x + e$

e is a random variable, which may have both positive and negative values, so

e is normally distributed $E(e) = 0$

the standard deviation of e, $s_{y\,x}$, is constant over the whole range of variation of x. This property is called “homoscedasticity.” since $E(e) = 0$, we’re supposing that

$E(y) = a + bx + E(e)$

$E(y) = a + bx$

Procedure for calculating simple linear regression

1. Compute the mean x and y, the corrected sum of squares, the cross product and the sum of cross products, and the corrected sum of cross products , of the variables x and y

2. Compute the estimates of the regression parameters α and β as α = y - β x (**α** and **β** are considered as the estimates of **a** and **b** rather than parameters).

Therefore β = $\frac{\sum xy}{\sum x^2}$

3. Then substitute the value of β in the linear equation:

$$\alpha = y - \beta x.$$

Thus the estimated linear regression is

$$y = \alpha + \beta x$$

4. Plot the observed points and draw a graphical representation of the estimated linear regression equation above.

5. Test the significance of β. First you calculate the residual mean square as:

$$\beta = S^2_{y,x} = \frac{\sum y^2 - {(\sum xy)^2}/{\sum x^2}}{n-2}$$

and compute the t_β value as:

$$t_\beta = \frac{\beta}{{S^2_{y,x}}/{\sum x^2}}$$

6. The Conclusion.

The regression coefficient β is said to be significantly different if the calculated t-value above is greater than the tabular value at the 5% and 1% levels of significance.

Scatter Diagram/Plot

A scatter diagram is a graph of the ordered pairs (x,y) of numbers consisting of the independent variable X and the dependent variable Y. It is a visual way to describe the nature of the relationship between x and y. i.e. it enable us to see whether there is any pattern among the points. The more distinct a pattern is, the more closely the two variables are related in some way.

For example: Construct a scatter diagram for the data.

Suppose that you had the following results from an experiment in which we measured the growth of goats at different Protein levels.

Ages, Month	Body wt, kg
0	10
1	20
2	25
3	32
4	33
5	35
6	47
7	49
8	53

You should explore your data by drawing a scattered graph. You can see that the weight of the goat is linearly related to the age of the goats in month. Regression analysis helps you to quantify this relationship.

	Ages, Month	Body wt, kg	x^2	x^2	xy
	0	10	0	100	0
	1	20	1	400	20
	2	25	4	625	50
	3	32	9	1024	96
	4	33	16	1089	132
	5	35	25	1225	175
	6	47	36	2209	282
	7	49	49	2401	343
	8	53	64	2809	424
Total	$\Sigma x = 36$	$\Sigma y=304$	$\Sigma x^2=204$	$\Sigma y^2=11882$	$\Sigma xy=1522$
	$\bar{x} = 4$	$\bar{y} = 33.778$	$s_x^2 = 7.5$	$S_y^2 = 201.69$	$cov_{xy} = 34$

The general equation for a fitted regression line is given as:

$Y = a + bX$

Where:

Y = Dependent variable on the vertical axis

X = Independent variable on the horizontal axis

a = Intercept

b = Slope of the regression line or the correlation coefficient of Y on X.

The intercept (a) is estimated as:

$a = y - b\,x$

Where

y = Mean of the dependent variable

x = Mean of the independent variable

$b = \frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2} = \frac{9(1522)-(36)(304)}{9(204)-36^2} = \frac{2754}{540} = 5.1$

Now estimate the intercept using the formula a = y – bx

In our example:

b = 5.1

y = 33.778

x = 4.00

a = 33.778 -5.1*4 = 13.378

So the linear equation will then be

Y = 13.378 + 5.1X

A dependent variable (Y) is the variable in regression that cannot be controlled or manipulated.

An independent variable (X) is the variable in regression that can be controlled or manipulated.

Example: An animal scientist is interested in studying the relationship between the age of goats and the liveweight of the goat. Body weight can be said to depend on the age. Here age is the independent variable (X) and body weight is the dependent variable (Y)

Types of Correlation and Regression

Correlation can be *Simple* or *Multiple*. In simple relationships, there are only two variables under study. **For example**, a researcher may wish to study the relationship between body length and weight in a population of chickens.

In *Multiple relationships* more than two variables are under study for example, an Animal Scientist may wish to investigate the relationship between number of eggs produce by laying chickens and factors such as different feed protein levels, quantity of feed given per day and hours of lighting per day.

Regression can be linear or curvilinear regression. Linear could be either simple linear regression or multiple linear regressions. Curvilinear – could be exponential, quadratic, and logarithmic etc.

Simple relationships can also be *Positive* or *Negative*. A positive relationship exists when the two variables under study increase or decrease at the same time. In a negative relationship as one variable increases the other variable decreases, and vice versa.

Generally, simple correlation and simple linear regression may be:

1. **Positive correlation** – when an increase in one variable is associated to a greater or lesser extent with an increase in the other.

2. **Negative correlation** – when an increase in one variable is associated to a greater or lesser extent with a decrease in the other.

3. **Perfect correlation** – when a change in one variable is exactly matched by a change in the other variable. If both increase together, it is perfect positive correlation: if one decreases as the other increases, it is perfect negative correlation.

4. **High correlation** – When a change in one variable is almost exactly matched by a change in the other.

5. **Low correlation** – when a change in one variable is to a small extent matched by a change in the other.

6. **Zero correlation** – when the two variables are not in matched at all, and there is no relationship between changes in one variable and changes in the other.

Spurious Correlation

When interpreting correlation, r, it is important to realize that, there may be no direct connection at all between highly correlated variables. When this is so, the correction is termed spurious or nonsense correlation. It can arise in two ways:

(a) There may be an indirect connection

(b) There may be a series of coincidences.

Covariance

When the association of two variables is assessed we can speak of the resulting assessment as the covariance ('Cov') of the variables. The use of analysis of covariance helps to eliminate variability. Covariance can be measured by finding the average of the products of the deviations of each of the paired variables from the overall mean of the relevant variable i.e.

Tutor marked assignment

1. Can you suggest 10 naturally occurring variable on your farm both crop and livestock that are related somehow?
2. Ten maize plants were randomly selected and treated weekly with a solution in which x g of Nitrogen fertilizer was dissolved in fixed quantity of water. The yield of maize grain was recorded. The resulting data is shown in the following table

	I	II	III	IV	V	VI	VII	VIII	IX	X
Ferterlizer, g	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5
yield, kg	4.0	5.5	6.8	8.0	8.5	8.7	9.5	9.8	9.5	9.9

a) Calculate the equation correlation coefficient;
b) Calculate the equation of the regression line
c) Estimate the yield of plant treated weekly with 3.3g of Nitrogen fertilizer
d) Give reason(s) why you cannot estimate the yield of plant treated weekly with 10 g of fertilizer.

3. Differentiate between correlation coefficient and regression equation.

a) Given the following information about buck weight and it weight at those ages in month

Ages, Month	Body wt, kg
0	10
1	20
2	25
3	32
4	33
5	35
6	47
7	49
8	53

b) Calculate the correlation between the ages and body weight of the animal
c) What is the regression coefficient?

d) Is it possible to predict weight of the animal at 20 months of age using your calculated regression equation? Give your reason

9.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

http://www1.appstate.edu/~mcraelt/simpreg1.pdf

Unit 10 Analysis of Covariance

10.1 Introduction

Sometime a variable that has not interested a researcher may seem to have effect on the outcome of an experiment. For example a researcher is interested in the effect of feeding three different types of forage to cattle. The response or dependent variable is the weight gain. The treatment variable is type of forage. The researcher is not interested in the initial body weight of the individual animal used in the experiment. But it is however noticed that the initial body weight is has effect on the weight gain. You then use the analysis of covariance ANCOVA to check if the regression of weight gain is the same for each of the type of forage. Ancova will tell you whether the regression lines are different from each other in either slope or intercept.

10.2 Objective

You will be able to know

Basic ideas behind ANCOVA

and define ANCOVA

When to use ANCOVA

Assumption of ANCOVA

State Null hypotheses for ANCOVA

How the test works

10.3 Main Text

10.3.1 Basic ideas behind ANCOVA

Covariates (concomitant variables) can reduce the MSE thereby increasing power testing. And as we have seen sometimes they are absolutely necessary in order to get accurate analysis.

A covariate can adjust for differences in characteristics of subjects in the treatment groups. Baseline or pretest values are often used as covariates

10.3.2 What is ANCOVA

It is Analysis of Variance with covariates. It can also be defined as a combination of ANOVA and regression.

10.3.3 When to use it

ANCOVA is used when you have some categorical factors and some quantitative predictors. The continuous variables are referred to as covariates or concomitant variables. It is similar to blocking. The concomitant variables are not necessarily of primary interest, but still their inclusion in the model will help explain more of the response, and hence reduce the error variance. In some cases, failure to include an important covariate can yield misleading results.

ANCOVA is used to account or adjust for "pre-existing" condition such as initial weight in live0stock experiments or soil moisture in agronomic experiments. Any baseline data that is of continuous variable can be adjusted for, using ANCOVA.

ANCOVA adjust the treatment mean to a common X when the randomization falls short (only when the treatment and X are independent). It reduces the magnitude of the error variance (s) just like blocking. This is done by explaining some of the unexplained variances (Residual Sum of Square) which then reduces the error variance in the model.

The use of ANCOVA also bring about greater experimental control. By controlling known extraneous variable, you gain greater insight into the effect of the predictor variable(s).

10.3.4 Assumptions

The covariate will not be in anyway related to the treatment variable (factor)

The covariate will be linearly related to the response and the relationship will be the same for all levels of the factor (No interaction between covariate and factor).

10.3.5 General (separate slopes) ANCOVA Model and Null Hypothesis for one-way ANCOVA

$Y_{ij} = \mu_i + \boldsymbol{t}_i + \boldsymbol{b}_i x_{ij} + e_{ij}$

As usual $e_{ij} \sim IID\ N(0, \sigma^2)$

You will set up a null hypothesis which is the probability that there is no effect or relationship.

The Null hypothesis should be written as follows:

There is no significant effect of independent variable on the dependent variable controlling for a covariate.

Consider the following example:

An animal scientist decided to test three type of forage plant on growing sheep. Which of the forage plant causes the fasted growth? The sheep used for this study were of different weight from at the beginning of the experiment. So the scientist will like to find out if there is still significant effect after controlling for the initial body weight of the experimental animals.

The null hypothesis:

There is no significant effect of forage plant type on the weight gain of sheep controlling for initial body weight.

The procedure for calculation in ANCOVA

fit two or more linear regressions of y against x (one for each level of the factor)

estimate different slopes and intercepts for each level

use model simplification (deletion tests) to eliminate unnecessary parameters

In livestock management, suppose you are modeling weight (the response variable) as a function of sex and age of the animals. Sex is a factor with 2 levels (male and female) and age is a continuous variable. The maximal model therefore has 4 parameters: two slopes (a slope for males and a slope for females) and two intercepts (one for males and one for females) like this:

$\text{weight}_{\text{male}} = a_{\text{male}} + b_{\text{male}} \times \text{age}$

$\text{weight}_{\text{female}} = a_{\text{female}} + b_{\text{female}} \times \text{age}$

= + male × weight a b age female = female + female

It is important to note that computer softwares now exists to aid in the computation of covariances.

Tutor marked assignment

1. Define in your own words ANCOVA
2. State the assumptions of ANCOVA
3. Why do you opt for ANCOVA instead of ANOVA?
4. Write out the mathematical model of one way ANCOVA.
5. Interpret all the terms in the model.

Fifteen goats were weighed before they were randomly allocated to four treatments of feeding trials. The following data were gathered.

goatID	Initial weight	Feed Type	final wt	wtg
1	12	1	18	6
4	10	1	15	5
7	10	1	15	5
10	13	1	18	5
13	10	1	17	7
2	11	2	18	7
5	9	2	16	7
8	10	2	18	8
11	15	2	22	7

14	9	2	15	6
3	16	3	24	8
6	13	3	20	7
9	15	3	22	7
12	16	3	24	8
15	13	3	21	8

Using your knowledge of ANCOVA, carry out a complete analysis of this goat data. Interpret your result. What is your conclusion?

10.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

http://www.biostathandbook.com/ancova.html

Module 3

Unit 11. Hypothesis testing of attribute data

11.1 Introduction

The chi-square test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. Do the number of individuals or objects that fall in each category differ significantly from the number you would expect? Is this difference between the expected and observed due to sampling error, or is it a real difference?

Chi-Square enables you to estimate whether a relationship exists, but how do you know how strongly the variables are related? Chi square tests will allow you to perform hypothesis testing on nominal and ordinal data.

11.2 Objective

You will

1. Know the requirement for chi square analysis
2. Know how to calculate expected cell counts under the null distribution in a contingency table.
3. Be able to determine the degrees of freedom for testing a null hypothesis against a specified alternative hypothesis.
4. Be able perform the Pearson and Likelihood Ratio Chi-Square tests.
5. Know the relationship between the Pearson Chi-Square test and the Z (Score)test.
6. Know how to identify the cells that contribute most strongly to a significant Chi-Square test.

7. Be able to tests the fit of the proportions in the obtained sample with the hypothesized proportions of the population

11.3 Main text

The chi-square test statistic is an overall measure of how close the observed frequencies are to the expected frequencies. It has the form

$\chi^2 =$

$$\text{sum}\left(\frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}\right)$$

The null hypothesis of independence is rejected if χ^2 is large, because this means that observed frequencies and expected frequencies are far apart. The chi-square curve is used to judge whether the calculated test statistic is large enough. We reject H_0 if the test statistic is large enough so that the area beyond it (under the chi-square curve with (r-1)(c-1) degrees of freedom) is less than .05.

Expected Frequencies

When you find the value for chi square, you determine whether the observed frequencies differ significantly from the expected frequencies. You find the expected frequencies for chi square in three ways:

I . You hypothesize that all the frequencies are equal in each category. For example, you might expect that half of the calves born in the National Animal Production Research Institute, Zaria are males and the other half will be females. You figure the expected frequency by dividing the number in the sample by the number of categories. In this example, where there are 100 new

calves born, and two categories male and female, you divide your sample of 100 by 2, the number of categories, to get 50 (expected frequencies) in each category.

2. You determine the expected frequencies on the basis of some prior knowledge. Someone told you that in Kaduna State 80% of the rural dwellers keeps poultry while 20% does not. You then administer a simple questionnaire randomly to a sample of 200 rural dwellers in Kaduna state. You will calculate your expected number or frequency of poultry farmers and non-poultry farmer using your expected 80% and 20% respectively. When you analyse you data you will expect 80% of 200 or 160 respondents to be poultry farmers while 20% or 40 respondents to be non poultry farmers.

You read a newspaper that in Kwara state, 55% of the farmers keeps goat, 20% keeps sheep, 15 keeps pigs and 20% keeps cattle. As a student of Agricultural Science you were asked to confirm whether this is true.

Procedure

You will have to set up your null hypothesis which states that there is no significant difference between the expected and observed frequencies.

The alternative hypothesis states they are different.

The level of significance (the point at which you can say with 95% confidence that the difference is NOT due to chance alone) is set at .05 (the standard for most science experiments.)

The chi-square formula used on these data is

X2 = (O - E)2 E is sum of df is the "degree of freedom" (n-1) X2 is Chi Squ

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O is the Observed Frequency in each category

E is the Expected Frequency in the corresponding category

χ^2 is the chi square

You are now ready to use the formula for χ^2 and find out if there is a significant difference between the observed and expected frequencies for the livestock farmers in Kwara State.

You will set up a worksheet; then you will follow the directions to form the columns and solve the formula.

Category	O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
Goat	95	80	15	225	2.813
Sheep	40	60	-20	400	6.667
Pig	30	20	10	100	5.000
Cattle	35	40	-5	25	0.625
				χ^2 =	15.104

2. After calculating the Chi Square value, find the "Degrees of Freedom."

This calculated as number of categories less 1

df = N – 1

= 4 – 1 = 3

3. Find the table value for Chi Square. Begin by finding the df found in step 2 along the left hand side of the table. Run your fingers across the proper row until you reach the predetermined level of significance (.05) at the column heading on the top of the table. The table value for Chi Square in the correct box of df and P=.05 level of significance is 7.815.

If the calculated chi-square value for the set of data you are analyzing (15.104) is equal to or greater than the table value (7.815), reject the null hypothesis. There Is a significant difference between the data sets that cannot be due to chance alone. If the number you calculate is LESS

than the number you find on the table, then you can probably say that any differences are due to chance alone.

In this situation, the rejection of the null hypothesis means that the differences between the expected frequencies (based upon what you read in the papers) and the observed frequencies (based upon the data you collected using your questionnaires) are not due to chance. That is, they are not due to chance variation in the sample you took; there is a real difference between them.

The steps in using the chi-square test may be summarized as follows:

1. Write the observed frequencies in column O Test Summary

2. Calculate the expected frequencies and write them in column E.

3. Use the formula to find the chi-square value:

4. Find the df (N-1)

5. Find the table value (consult the Chi Square Table using the df and $\alpha = 0.05$)

6. If your chi-square value is equal to or greater than the table value, reject the null hypothesis: differences in your data are not due to chance alone

For example, the reason observed frequencies in a fruit fly genetic breeding lab did not match expected frequencies could be due to such influences as:

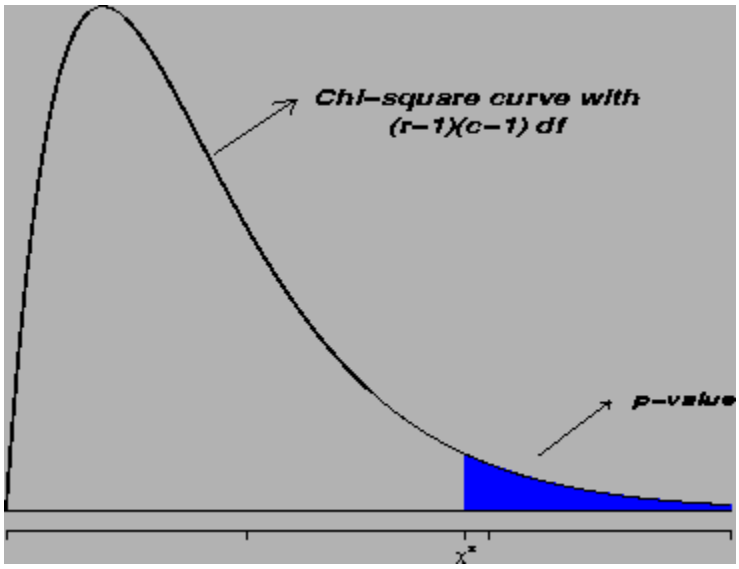
• Mate selection (certain flies may prefer certain mates)

• Too small of a sample size was used

• Incorrect identification of male or female flies

• The wrong genetic cross was sent from the lab

• The flies were mixed in the bottle (carrying unexpected alleles)



Requirement for Chi Square analysis

Chi-Square Test Requirements are as listed

1. Quantitative data. 2. One or more categories. 3. Independent observations. 4. Adequate sample size (at least 10). 5. Simple random sample. 6. Data in frequency form. 7. All observations must be used.

10.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

Unit 12. Goodness of fit

12.1 Introduction

The chi square is particularly useful for analyzing ordinal and nominal data. It is called non-parametric analysis. It is widely used in agriculture. In this unit, you we learn the use of chi-square to test for goodness of fit.

12.2 Objectives

After studying this unit, you should be able to

Formulate null and alternative hypotheses for goodness of fit analysis

Calculate expected frequencies for a variety of probability models

Use χ^2 distribution to test if a set of observations fits an appropriate probability model

Uses sample data to test hypotheses about the shape or proportions of a population distribution

Tests the fit of the proportions in the obtained sample with the hypothesized proportions of the population

Evaluate effect size using phi coefficient or Cramer's V

Explain when chi-square test is appropriate

Test hypothesis about shape of distribution using chi-square goodness of fit

Test hypothesis about relationship of variables using chi-square test of independence

12.3 Main Text

12.3.1 Goodness of Fit Null Hypothesis

Null hypothesis: This is the statement that should not be rejected when the discrepancy between the Observed and Expected values is small. It should be rejected when the discrepancy between the Observed and Expected values is large.

Chi-Square distribution includes values for all possible random samples when H_0 is true

All chi-square values are greater than or equal to 0. When H_0 is true, sample χ^2 values should be small.

Rationale for null hypotheses:

No preference (equal proportions) among categories, *OR*

No difference in specified population from the proportions in another known population

Goodness of Fit Alternative Hypothesis

Often equivalent to "…population proportions are not equal to the values specified in the null hypothesis…"

12.3.2 Goodness of Fit Test Data

Individuals are classified (counted) in each category, e.g., grades; exercise frequency; etc.

Observed Frequency is tabulated for each measurement category (classification)

Each individual is counted in one and only one category (classification)

Expected Frequencies in the Goodness of Fit Test

Goodness of Fit test compares the Observed Frequencies from the data with the Expected Frequencies predicted by null hypothesis.

You will construct Expected Frequencies that are in perfect agreement with the null hypothesis Expected Frequency is the frequency value that is predicted from H_0 and the sample size; it represents an idealized sample distribution

12.3.3 Chi-Square Degrees of Freedom

Chi-square distribution is positively skewed and belong to a family of distributions determined by the associated degrees of freedom.

The chi square distribution curves assumes slightly different shape for each value of degrees of freedom. Degrees of freedom for Goodness of Fit Test is calculated using the following formula:

$Df = C – 1$

Where df is degrees of freedom and

C is the number of categories

12.3.4 Locating the Chi-Square Distribution Critical Region

In order to locate the critical region of the Chi-Square you will have to do the following:

- Determine significance level (alpha). This could be any value less than 1. In agricultural research, 0.05 is used.
- Locate critical value of chi-square in a table of critical values according to
    - Value for degrees of freedom (df)
    - Significance level chosen

12.3.5 How to report your results

Your report should be presented properly describing whether there were significant differences between category preferences or not. Your report should include χ^2, df, sample size (n) and test statistic value and significance level

For example $\chi^2 (3, n= 50) = 8.08, p< .05$

Tutor marked assignment

1 The expected ratio of white, black and multicolour rabbits in a population are 0.34, 0.52 and 0.14 respectively. In a sample of 800 rabbits there were 280 white, 480 brown and 40 multicolour. Are the proportions in that sample of rabbits different than expected?

2 The past study showed the proportions of farmer's land used for growing certain crops in crops in Kaduna State

	Crop	Proportion of land used for
1	maize	0.4
2	millet	0.3
3	guinea corn	0.2
4	soyabean	0.1

Ten years after the study, a student suspect that the situation might have changed. So he conducted a study to find the proportion of farmer's land used for the crop in the last cropping season. The result is hereby summarized

	Crop	Observed land used in ha
1	maize	75
2	millet	40
3	guinea corn	20
4	soyabean	10

Is the new sample proportion different from the expected?

12.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

Unit 13 Chi-Square Test for Independence

13.1 Introduction

In the last unit, you learnt how to do test of goodness of fit using Chi-Square analysis. Another use of Chi-Square is the test of independence among 2 ordinal or nominal variables.

13.2 Objective

You will be able to

1 Identify the type of data and arrange the data in matrix form

2 Formulate Null Hypothesis and its alternative for Test of Independence

3 Differentiate between observed and expected frequencies

4 Computing Expected Frequencies

13.3 Main text

13.3.1 Arrangement of data

Chi-Square Statistic can test for evidence of a relationship between two nominal or ordinal variables. You must arrange your data in a matrix format. The counts or frequencies are presented in the cells of the matrix. Each is individual jointly classified on each variable. The design may be experimental or observational.

Frequency data from a sample is used to test the evidence of a relationship between the two variables in the population using a two-dimensional frequency distribution matrix

13.3.2 Null Hypothesis for Test of Independence

Null hypothesis: the two variables are independent (no relationship exists)

Two versions

Single population: No relationship between two variables in this population.

Two separate populations: No difference between distributions of variable in the two populations

Variables are independent if there is no consistent predictable relationship

13.3.3 Observed and Expected Frequencies

Frequencies in the sample are the Observed frequencies for the test

Expected frequencies are based on the null hypothesis prediction of the same proportions in each category (population)

Expected frequency of any cell is jointly determined by its column proportion and its row proportion

13.3.4 Computing Expected Frequencies

To determine the expected frequency for each cell in the matrix or contingency table, under the assumption that the two variables are independent, you will use the following equation

$$E_{r,c} = \frac{\text{Total row} \times \text{Total of column}}{\text{Total Number of Observations}}$$

Where $E_{r,c}$ = the expected frequency of the cell that corresponds to the intersection of Row r and Column c

13.3.5 Computing Chi-Square Statistic for Test of Independence

The computation is similar to that of the Chi-Square Test of Goodness of Fit

Chi-Square Statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Degrees of freedom (df)= (R-1)(C-1)

Where R is the number of rows and C is the number of columns

13.3.6 Measuring Effect Size for Chi-Square

A significant Chi-square hypothesis test shows that the difference did not occur by chance but does not indicate the size of the effect

For a 2x2 matrix, the phi coefficient (Φ) measures the strength of the relationship

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

So φ^2 would provide proportion of variance accounted for just like r^2

Effect size in a larger matrix

For a larger matrix, a modification of the phi-coefficient is used: Cramer's V

$$V = \sqrt{\frac{\chi^2}{n(df^*)}}$$

df^* is the smaller of (R-1) or (C-1)

13.3.7 Interpreting Cramer's V

For df* =	Effect		
	Small	Medium	Large
1	0.10	0.30	0.50
2	0.07	0.21	0.35
3	0.06	0.17	0.29

Tutor marked Assignment

1. How do females and males compare in the pursuit of Agricultural course? The table below present counts (in thousands) from a faculty of Agriculture admission list categorized by the course admitted for and the sex of the student.

	Agric Econs	Animal Sci	Agric Extension	Soil Sci
Female	642	227	32	18
Male	522	179	45	27

Perform a Chi-square test of homogeneity. Use a 1% significance level.

2. A student read in the farmer's magazine that religion is a determinant of the species livestock kept by farmers in Kaduna state. He decided to interview random samples of farmer in a village in Kaduna state. The result is tabulated as follows:

	Sheep	Cattle	Pigs
Moslem	72	50	10
Christian	58	30	40
Traditionalist	15	10	5

Determine if the type of livestock being kept in this community actually depend on the religion of the farmer. Use $\alpha = 0.05$ to make your conclusion

13.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.

Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.

Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)

Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company

Unit 14. Field experimentation, collection and processing of data

14.1 Introduction

In agricultural and indeed biological experiments, the research have to go to the field to collect that for analysis that will eventually be used to support his hypothesis. The field could be an open land which is being used to grow some kind of crops. It could be the laboratory where specific chemical analysis is being carried out. It could even mean administration of questionnaire to households in some specific villages.

14.2 Objectives

You will be able to

a) Learn the essentials of experimentation

b) Understand the principles of field experimentation

c) Handle experimental material

14.3 Main Text

14.3.1 Essential in Steps Experimentation

The correct choice of a statistical procedure for any experiment must be based on sound knowledge of statistics and the subject matter of the research. Thus, a good experiment can be designed by:

1. A subject matter specialist (SMS) with some training in experimental statistics
2. A statistician with some background and experience in the subject matter under experimentation
3. A joint effort and cooperation between a statistician the SMS Problem definition

You begin by asking questions from what you have observed over the years personally or in literature

Establish if there is truly a problem

After identifying the problem, there is need to do a comprehensive literature review to know how far other scientists have gone on the subject matter and then identify the missing gaps in knowledge. This procedure is very important because it avoids duplication of efforts.

Definitions of Objectives

There would be one or more objectives in a research study. The objective should be very brief or concise, self explanatory and achievable. Example :

Objective: To evaluate the yield potential of new soyabean varieties.

You can have broad objective or general objective and (specific objectives) which always address the specific topic. Avoid rewriting your project title as your objective e.g.

Project title: Study of the effect of fertilizer on maize varieties cassava and rewriting that as objective is wrong. The objective in this case is to determine the effect of fertilizer on yield and yield component of maize.

Choice of Treatments: Before a researcher can define experimental treatment, a very good knowledge of the subject matter is needed. An experimental treatment is any process/procedure whose effect is to be measured and compared with others.

The quantitative or qualitative component of a treatment is called the treatment levels. When there are two or more types of treatments, then each treatment is referred to as factor. For example, in the study of the effects of insecticide on the control of insect pests on tomato varieties, there are **two factors** here, namely: the **insecticides** to be used and the number of **tomato varieties** in the study.

Definition of Experimental Material Any object or element of the environment on which treatment is applied and observation made is called experimental material. The portion of unit of experimental material receiving the application of treatment is called the experimental unit or experimental plot in a field experiment. The minimum requirements for a valid experimental design are:

Replication: It is the application of a treatment more than once in an experiment. This is done

1. To provide an estimate of experimental errors 2.
2. To improve the precision of an experiment
3. To increase the scope of inference
4. To reduce or control error variance Randomization: Every treatment must be given equal chance of being placed within and experimental plot or a sample unit.

Randomization of treatments indicates that the experimental error is randomly distributed, that is the residual error is not clustered. Random distribution of error is important in order to measure the level of statistical significant of effects of treatments or factors. Randomization is done to avoid bias in the allocation of treatments to plots.

Blocking: This is accomplished by blocking or subdividing the experimental area or experimental material to more or less homogenous groups.

14.3.2 Principles of field experimentation

An experiment is a planned investigation that is carried out to obtain additional knowledge in order to solve identified problems and to obtain solutions to the problems.

Identify the problem

The problems can be identified from a survey, personal experience and literature search. The identified problem must be state in unambiguous term. –

Literature review

After identifying the problem, the next step is to carry out literature review – to find out how other scientists in other locations or countries had tried to solve the problem. What experimental procedure they used and the results obtained in order to ensure that your proposed experiment is properly conducted. Literature review is made easy by electronic search, library, e-mail correspondence etc. –

Clearly state the objective of the work.

The objective must be straight forward and simple. It must be specified to ensure that the study is properly focused and the right results are obtained.

Setting up of hypothesis:

The hypothesis is state in the negative. This null hypothesis (Ho) states for instance that there are no differences in the yield of the varieties to be evaluated. The alternate hypothesis (Ho), will be accepted if the experiment shows otherwise.

Designing of experiment:

To be able to answer the problem, the scientist will - Conduct the experiment

Collect data

Statistically analyse the data - Interpret the data and Report the results obtained

Field experimentation practices

Field preparation

Choice of treatment and factors

Choice of design and number of replications

Plot labeling

Treatment randomization and layout of the experiment

Planting and application of treatments in the field

Handling of experimental material

Experimental unit and sampling unit

Size of guard or border row

Data collection, handling and processing

Filed preparation must be timely including ploughing, harrowing, leveling and removing the entire stump. Choice of design depends on the location, homogeneity of the site. It depend on the number of factors to be investigated. Treatment randomization is very important to avoid bias in allocation of treatment to experimental plots. Randomization is important to correct and minimize residual error. The smaller the experimental error, the better is the precision of the experiment. Field layout: In the field lay out the site, location, orientation relation to N.S.E.W. or road, the number of plot, replicate, the treatment number must be clearly shown. It must be typed and printed and given to all participants in the experiment. Planting and application of treatments in the field has to be done at the right time. The treatment number must be put on label and place on the plot.

14.3.3 Handling of experimental material

Experimental materials must be carefully handled to avoid spoilage, leakage, overdose or under dose during measurement, transportation and application on the field. Also, contamination of contaminable materials must be avoided.

Experimental unit and sampling unit: The experimental unit is usually called the Gross plot size. It is the smallest unit in the experiment. Each experimental unit receives a treatment. The sampling unit is a portion of the experimental unit. It defines the portion from where sample is taken for measurement.

Size of guard/border row: The size of an experimental unit or plot must be large enough to avoid border effect. That is there must be guards rows.

The advantages of border row include:

1. It gives protection or shield or break when applying different fertilizer rates or pesticides from one plot to the other plot.

2. It helps to reduce border effect because border plants grow better than other plants within the plot.

Data collection

Collect the data in a log book

Summarize data the same day that you collect the data

Avoid using loose paper to collect data

It must be readable by other persons

Where sample has to be weighed or dried, it should be labeled and properly packed in a paper bag. - Counting and weighing procedure must be adequate - Summarized data must carry sampling date, name of variable, unit of measurement and location of the experiment - The summarized data must be neatly presented for statistical analysis

Tutor marked assignment

1. Give 2 reason why a research must do a literature search before conducting a study
2. The subject matter specialists must be involved in the design of experiment. Which other professional is critical to good experimental design?
3. List 4 reasons for replication in experimental design

4.5 Further Reading

Devore, J. L., & Peck, R. (1986). *Statistics: The exploration and analysis of data*. St. Paul: West Pub. Co.

Donnelly R. A. (2004). *The complete idiot's guide to statistics* (Vol. The complete idiot's guide). Indianapolis, IN: Alpha.
Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research (2nd ed.). New York: Wiley.
Kaps, M., Lamberson, W. R., & Lamberson, W. (2004). **Biostatistics for animal science**. Wallingford: CABI Publishing.
Jaisingh, L. R. (2006). **Statistics** for the **utterly confused** (2nd ed.). New York: McGraw-Hill.
McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland (http://www.biostathandbook.com/index.html)
Weiss N.A. 1999. *Elementary Statistics, fourth edition*. Reading, Massachusetts: Addison-Wesley Publishing Company