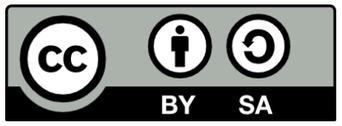
STA 203: STATISTICS FOR PHYSICAL SCIENCES AND ENGINEERING

Published by the Centre for Open and Distance Learning,
University of Ilorin, Nigeria

E-mail: codl@unilorin.edu.ng
Website: https://codl.unilorin.edu.ng



Course Development Team

Content Authoring

Abidoye, Adekunle Omotayo
B.Sc.,Statistics (Ilorin), M.Sc. Statistics (Ibadan) Ph.D. Statistics (Ilorin).

Content Editor

Mahmud Abdulwahab
Center for Open and Distance (CODL)
University of Ilorin, Nigeria

Instructional Design

Mr. Olawale S. Koledafe
Center for Open and Distance (CODL)
University of Ilorin, Nigeria

Miss Damilola Adesodun
Department of Educational Technology,
University of Ilorin, Nigeria

Miss Kareem Haleemat
Department of Educational Technology,
University of Ilorin, Nigeria

Sanusi Ridwan Opeyemi
Department of Educational Technology,
University of Ilorin, Nigeria

From the Vice Chancellor

Courseware development for instructional use by the Centre for Open and Distance Learning (CODL) has been achieved through the dedication of authors and the team involved in quality assurance based on the core values of the University of Ilorin. The availability, relevance and use of the courseware cannot be timelier than now that the whole world has to bring online education to the front burner. A necessary equipping for addressing some of the weaknesses of regular classroom teaching and learning has thus been achieved in this effort.

This basic course material is available in different electronic modes to ease access and use for the students. They are available on the University's website for download to students and others who have interest in learning from the contents. This is UNILORIN CODL's way of extending knowledge and promoting skills acquisition as open source to those who are interested. As expected, graduates of the University of Ilorin are equipped with requisite skills and competencies for excellence in life. That same expectation applies to all users of these learning materials.

Needless to say, that availability and delivery of the courseware to achieve expected CODL goals are of essence. Ultimate attention is paid to quality and excellence in these complementary processes of teaching and learning. Students are confident that they have the best available to them in every sense.

It is hoped that students will make the best use of these valuable course materials.

Professor S. A. Abdulkareem
Vice Chancellor

Foreword

Courseware remains the nerve centre of Open and Distance Learning. Whereas some institutions and tutors depend entirely on Open Educational Resources (OER), CODL at the University of Ilorin considers it necessary to develop its own materials. Rich as OERs are and widely as they are deployed for supporting online education, adding to them in content and quality by individuals and institutions guarantees progress. Doing it in-house as we have done at the University of Ilorin has brought the best out of the Course Development Team across Faculties in the University. Credit must be given to the team for prompt completion and delivery of assigned tasks in spite of their very busy schedules.

The development of the courseware is similar in many ways to the experience of a pregnant woman eagerly looking forward to the D-day when she will put to bed. It is customary that families waiting for the arrival of a new baby usually do so with high hopes. This is the apt description of the eagerness of the University of Ilorin in seeing that the centre for open and distance learning [CODL] takes off.

The Vice-Chancellor, Prof. Sulyman Age Abdulkareem, deserves every accolade for committing huge financial and material resources to the centre. This commitment, no doubt, boosted the efforts of the team. Careful attention to quality standards, ODL compliance and UNILORIN CODL House Style brought the best out from the course development team. Responses to quality assurance with respect to writing, subject matter content, language and instructional design by authors, reviewers, editors and designers, though painstaking, have yielded the course materials now made available primarily to CODL students as open resources.

Aiming at a parity of standards and esteem with regular university programmes is usually an expectation from students on open and distance education programmes. The reason being that stakeholders hold the view that graduates of face-to-face teaching and learning are superior to those exposed to online education. CODL has the dual-mode mandate. This implies a combination of face-to-face with open and distance education. It is in the light of this that our centre has developed its courseware to combine the strength of both modes to bring out the best from the students. CODL students, other categories of students of the University of Ilorin and similar institutions will find the courseware to be their most dependable companion for the acquisition of knowledge, skills and competences in their respective courses and programmes.

Activities, assessments, assignments, exercises, reports, discussions and projects amongst others at various points in the courseware are targeted at achieving the objectives of teaching and learning. The courseware is interactive and directly points the attention of students and users to key issues helpful to their particular learning. Students' understanding has been viewed as a necessary ingredient at every point. Each course has also been broken into modules and their component units in sequential order.

At this juncture, I must commend past directors of this great centre for their painstaking efforts at ensuring that it sees the light of the day. Prof. M. O. Yusuf, Prof. A. A. Fajonyomi and Prof. H. O. Owolabi shall always be remembered for doing their best during their respective tenures. May God continually be pleased with them, Aameen.

Bashiru, A. Omipidan
Director, CODL

Introduction

As introduction, we shall look at the measures of location or central tendency in detail, we also consider the various measures of central tendency such as Arithmetic mean, Median, Mode, Geometric and Harmonic mean as well. We let the students know the advantages and disadvantages of the measures of location or central tendency. We also discussed the measures of dispersion such as Range, variance, Semi Interquartile range and coefficient of variation. In measures of location, we consider the grouped and ungrouped data for the students and teach them how they calculate grouped and ungrouped data. The course introduces the students to the concepts of probability: experiment, conditional, outcome, trial, chance, elementary event, event, empty set and sample space, Expectations, variance, distribution functions; Probability distributions: Bernoulli, Binomial, Poisson, Geometric, Uniform and Normal distributions. Regression: concept of Simple linear equation, Analysis of variance for regression; Correlation: Pearson correlation.

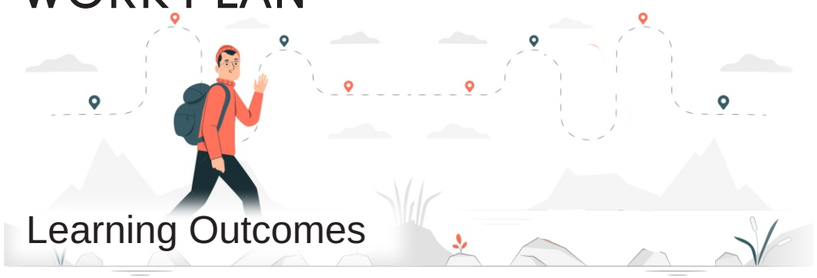
Course Goal

The course is designed to introduce students in the Physical Sciences and Engineering to the application of statistics in their discipline. This course is to intimate the students with the usefulness of statistics in their various field of studies. This course is to enlighten the student on the importance of statistics in carrying out researches in their various areas of study. This course is also to develop in students the ability to apply their knowledge and skills to the solution of theoretical and practical problems in Statistics.

Related Courses: NIL

Prerequisite: NIL

WORK PLAN



Learning Outcomes

Required For

This is a compulsory course for students in Departments of Chemistry, Industrial Chemistry, Geology, Mathematics, Computer Science. Students are expected to participate in all the course activities and have minimum of 75% attendance to be able to write the final examination

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

i Use different scales of measurement,

ii Construct use of frequency distributions,

iii. Differentiate between different measures of location and when to use them,

iv. Recognize different measures of spread.

v. Illustrate the concepts of probability.

vi. Compare some discrete and continuous probability distributions and their different applications in real life situations.

vii. Fit a simple linear Regression model,

viii. Express the concepts of correlation

Week 01

Course Guide

Module 1

Introduction of the measures of location or central tendency

Unit 1: Introduction of the measures of location or central tendency

Unit 2: Grouped and Ungrouped data

Unit 3: Measures of dispersion or Measures variability

Week 02

Module 2

Introduction to Probability

Unit 1 : Introduction to Probability

Unit 2 : Experimental Probability

Unit 3 : Law of Probability

Module 3

Binomial distributions

Unit 1 : Binomial distributions

Unit 2 : Continues Binomial distributions

Week 03

Module 4

Poisson distributions

Unit 1 : Poisson distributions

Unit 2 : Continues Poisson distributions

Module 5

Normal distributions

Unit 1 : Normal distributions

Unit 2: Linear Regression

Week 04

Module 6

Analysis of Variance for Simple Linear Regression

Unit 1 : Analysis of Variance for Simple Linear Regression

Unit 2 : Correlation

Unit 3: Continues Correlation

Course Requirements

Requirements for success

The CODL Programme is designed for learners who are absent from the lecturer in time and space. Therefore, you should refer to your Student Handbook, available on the website and in hard copy form, to get information on the procedure of distance/e-learning. You can contact the CODL helpdesk which is available 24/7 for every of your enquiry.

Visit CODL virtual classroom on http://codllms.unilorin.edu.ng. Then, log in with your credentials and click on STA 203. Download and read through the unit of instruction for each week before the scheduled time of interaction with the course tutor/facilitator. You should also download and watch the relevant video and listen to the podcast so that you will understand and follow the course facilitator.

At the scheduled time, you are expected to log in to the classroom for interaction.

Self-assessment component of the courseware is available as exercises to help you learn and master the content you have gone through.

You are to answer the Tutor Marked Assignment (TMA) for each unit and submit for assessment.

Embedded Support Devices

Support menus for guide and references

Throughout your interaction with this course material, you will notice some set of icons used for easier navigation of this course materials. We advise that you familiarize yourself with each of these icons as they will help you in no small ways in achieving success and easy completion of this course. Find in the table below, the complete icon set and their meaning.

 **Introduction**	 **Learning Outcomes**	 **Main Content**

 **Summary**	 **Tutor Marked Assignment**	 **Self Assessment**
 **Web Resources**	 **Downloadable Resources**	 **Discuss with Colleagues**
 **References**	 **Futher Reading**	 **Self Exploration**

Grading and Assessment

TMA

CA

Exam

Total

Picture: 01

Photo:Vectorstock.com

Module 1

Introduction Of The Measures Of Location Or Central Tendency

Picture: 02

Photo:Unsplash.com

UNIT 1

Introduction of the measures of location or central tendency.

Introduction

Measures of location or Central Tendency: The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of a data set are the mean (average) and the median. To calculate the mean weight of 50 people, add the 50 weights together and divide by 50. To find the median weight of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

Learning Outcomes

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

1. Discuss the importance of the measures of location
2. Compute the various measures of location such as mean, mode and median
3. Differentiate the measures of location for the grouped and ungrouped data

Main Content

4 mins

Introduction of the measures of location or central tendency

Our main focus on this chapter is to develop measures that can be used to summarize a data set. These measures formally called statistics are quantities whose values are determined by the data. We study the sample mean, sample median and sample mode. These are all statistics that measure the centre or middle value of a data set. Statistics that indicate the amount of variation in the data set are also considered. We learn about what it means for a data set to be normal and we present an empirical rule concerning such sets.

Compute the measure of location with the following set of data

Mean: is the set of numbers; $\mathbf{x_1, x_2 \ldots x_n}$ their arithmetic mean is defined as their sum divided by **n.** The advantage of is mean is very easy to calculate, it take into consideration all members of the data set, it is widely used, we can obtain the mean of two or more means of set of data and it is very reliable.

Example

(1) The average fuel efficiencies in miles per gallon of cars sold in the United States in the years 1999 to 2003 were: 28.2, 28.3, 28.4, 28.5, and 29.0. Find the sample mean of this set of data.

Solution
snisn

The samples mean which we designate by *x* is defined by

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

The sample mean is the average of the five data values. Thus,

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{142.4}{5} = 28.48$$

The Median: Here, you will order the data values from smallest to largest. If the number of data values is odd, then the sample median is the middle value in the ordered list; if it is even, then the sample median is the average of the two middle values. It follows from this definition that if there are three data values, then the sample median is the second-smallest value and if there are four, then it is the average of the second and third smallest values. The advantages of median are it always exists, that is it can be found for any set of numerical data, it is always unique, it is not easily affected by the extreme values and the median can be obtained even if the distribution is open at either end. Illustrate the following data represent the number of weeks it took seven individuals to obtain their driver's licenses.

For example:

Find the sample median.
2, 110, 5, 7, 6, 7, 3.

Solution
You should note that when you are finding the median you will have to first arrange the data in increasing order.
i.e. 2, 3, 5, 6, 7, 7, 110

So, since the sample size is 7, it follows that the sample median is the fourth smallest value.

That is, the sample median number of weeks it took to obtain a driver's license median is M= 6 weeks.

The Mode

Sample Mode is another indicator of central tendency, which is the data value that occurs most frequently in the data set. In other word, the number that appears most frequently is called the mode.

Example

The following are the sizes of the last 8 dresses sold at a women's boutique. What is the sample mode?

8, 10, 6, 4, 10, 12, 14, 10.

Solution

The sample mode is 10, since the value of 10 occurs most frequently.

If no single value occurs most frequently, then all the values that occur at the highest frequency are called modal values. In such a situation we say that there is no unique value of the sample mode.

Summary

You have learnt how to calculate the value of mean, mode and median with given data example were also given for students to understand.

Self Assessment Questions

(1) Calculate the mean weight measure in kilogram, of eight malnourished adults placed on a special diet; whose weights are 45, 40, 50, 43, 49, 47, 46, 41

(2) Obtain the mode of these following observations 8, 9, 2,4, 8, 9, 15, 9

(3) The intelligence quotients (IQ's) of 10 boys in a class are given below: 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Find the mean I.Q.

(4) Calculate the mean value of the following observations 12, 19, 21, 30, 13, 19, 22, 31, 17, 20, 24, 31, 18, 21, 27, 31.

(5) Calculate mean value from the following data of the heights in inches of a group of students: 61, 62, 63, 61, 63, 64, 64, 60, 65, 63, 64, 65, 66, 64

(6) Find the median of the following set of dat 38, 34, 39, 35, 32, 31, 37, 30, 41

(7) Find the median of the following set of data 30, 31, 36, 33, 29, 28, 35, 36

(8) The following data were collected by a survey group.
12, 19, 21, 30, 13, 19, 22, 31, 17, 20, 24, 31, 18, 21, 27, 31
Compute the arithmetic mean of the observations.

(9) The following data were collected by a survey group.
12, 19, 21, 30, 13, 19, 22, 31, 17, 20, 24, 31, 18, 21, 27, 31
Compute the mode using the sixteen observations given.

(10) Find the median of the following numbers: 28, 29, 39, 38, 33, 37, 26, 20, 15, 25

Tutor Marked Assignment

- Obtain the mode of the numbers; 8, 10, 9, 9, 10, 8, 11, 8, 10, 9, 8, 8, 14
- Find the mode of the following distribution: 7, 4, 3, 5, 6, 3, 3, 2, 4, 3, 4, 3, 3, 4, 4, 2, 3
- Find the median of the following numbers: 2.64, 2.50, 2.72, 2.91, 2.35
- AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47
- Calculate the mean and the median.
The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.
3; 4; 5; 7; 7; 7; 7; 8; 8; 9; 9; 10; 10; 10; 10; 10; 11;
12; 12; 13; 14; 14; 15; 15; 17; 17; 18; 19; 19; 19; 21 ; 21; 22; 22
23; 24; 24; 24; 24
- Suppose that in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center": the mean or the median?

- Statistics exam scores for 20 students are as follows:
  50; 53; 59; 59; 63; 63; 72; 72; 72; 72; 72; 76; 78; 81; 83; 84; 84; 84; 90; 93
  Find the mode.

- Calculate the mean of the following sample: 1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

- Calculate the median of the following sample: 1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

References

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.
- Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.
- Gupta (2014). Introduction to Fundamental Statistics.
- Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.
- Introductory Statistics 3rd Edition by Sheldom M. Ross

Ungrouped data	Grouped data			
	Discrete		continuos	
	x	f	x	f
21,22,23,27	21	11	20-30	11
	22	7		
	27	19		

Picture: 03

Photo:Wikepedia.com

UNIT 2

Grouped and Ungrouped data

Introduction

It is very important that the numerical findings of any study be presented clearly and concisely and in a manner that enables one to quickly obtain a feel for the essential characteristics of the data. This is particularly needed when the set of data is large, as is frequently the case in surveys or controlled experiments. Indeed, an effective presentation of the data often quickly reveals important features such as their range, degree of symmetry, how concentrated or spread out they are, where they are concentrated, and so on. In this chapter we will be concerned with techniques, both tabular and graphic, for presenting data sets.

Our description of any object would have to depend partly on the nature of the object itself and partly on the purpose we might have for giving the description. The same argument holds also to the description of numerical data. The type of description we may choose or the statistical techniques we may employ also depend partly on the nature of the data themselves and partly on the purpose that we may have in mind.

Learning Outcomes

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

1. Calculate the grouped data for the grouped mean.
2. Calculate the grouped data for the grouped median.
3. Calculate the grouped data for the grouped mode.
4. Calculate upper quartile, 4th decile and 25th percentile

Main Content

How To Obtain Data For Grouped Data To Calculate The Mean, Mode And Median

5 mins

Do not forget that, once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. This is done by creating a table of frequencies which is known as a frequency distribution. There are three frequency distributions to choose from, depending on if the data is categorical or quantitative. When it is quantitative data, there are two frequency distributions to choose from. The three quartiles of a distribution are defined as values which divide it into four parts containing equal numbers of observations such as first quartile, second quartile and third quartile. Also we have deciles are values which by definition, divide a distribution or the area of the rectangles of its histogram into 10 equal parts. We shall write D_1 for the first decile which exceeds the lowest 10 percent of the data, D_2 for the second decile which exceeds the lowest 20 percent of the data, and in general D_i for the i^{th} decile which exceeds the lowest i 10 percent of the data. When calculating the frequency, you may need to round your answers so that they are as precise as possible.

The Grouped Mean

It will interest you to know that this first type of frequency distribution is also used when there is quantitative data. However, it is used when the range is large and the data values need to be grouped together. For example, 28 students were asked how many hours they worked per week. Their responses, in hours, are as follows: 15; 26; 13; 33; 22; 14; 27; 15; 32; 23; 5; 26; 25; 14; 34; 13; 15; 22; 15; 28; 10; 18; 21; 24; 20; 18; 34; 20; Here there are too many different data values to list them separately as in the ungrouped frequency distribution. Notice the range is 29 (highest – lowest = 34 – 5). Therefore we need to construct a grouped frequency distribution to group data values into classes. A class is an interval where the lowest value of the interval is known as the lower limit and the highest value of the interval is known as the upper limit.

Guidelines for classes:

i There should be between 5 and 20 classes
ii Classes must be mutually exclusive (no overlap of data values)
iii Classes must be all inclusive and continuous
iv Classes must be equal in width 42

Constructing a Grouped Frequency Distribution:

The following are steps involved in constructing a grouped frequency distribution.

(1) The first step is to find the Range (i.e. the highest data value – lowest data value)
(2) Then you determine the number of classes (usually the minimum is 5 classes and a maximum of 20 classes)
(3) Find the Class Width = Range and number of classes
NOTE: Round up
(4) Choose first lower limit (usually the lowest data value)
(5) Create the other lower limits of the classes by adding the class width to the previous lower limit
(6) And lastly create the upper limits by not overlapping the limits

Illustration on Grouped data:

Estimating the Mean from Grouped Data and we have the following data:

Seconds	Frequency
51 – 55	2
56 – 60	7
61 – 65	8
66 – 70	4

Obtain the mean, Median and Mode

The groups (51-55, 56-60, etc), also called class intervals, are of width 5

We can estimate the Mean by using the midpoints.

The midpoints are in the middle of each class: 53, 58, 63 and 68

Let's now make the table using midpoints:

Midpoint	Frequency
53	2
58	7
63	8
68	4

The quick way to do it is to multiply each midpoint by each frequency and see the table below

Midpoint x	Frequency f	Midpoint × Frequency fx
53	2	106
58	7	406
63	8	504
68	4	272
Totals:	**21**	**1288**

Then we add or sum all frequency up and divide by 21.

And then our estimate of the mean time to complete the race is:

$$\textbf{Mean} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1288}{21} = 61.333$$

(2) The median

The formula to calculate the Median is defined below

$$\textbf{Median} = L_1 + \left(\frac{N}{2} - f_{bm} f_m\right)C$$

where:

L_1 is the lower class boundary of the group containing the median

N is the total number of values

bmf is the cumulative frequency of the groups before the median group

mf is the frequency of the median group

C is the group width or constant

Computation

$L_1 = 60.5$, N = 21, *bmf* = 9, *mf* = 8, C = 5

Median = 60.5 + *(21/2) – 9* **8** × 5

= 60.5 + 0.9375

= 61.4375

The formula to calculate the Mode is defined below:

$$\text{Mode} = \left(L_1 + \left(\frac{f_m - f_{bm-1}}{(f_m - f_{bm-1}) + (f_m - f_{bm+1})}\right)\right)C$$

where:

L_1 is the lower class boundary of the modal group

f_{bm-1} is the frequency of the group before the modal group

f_m is the frequency of the modal group

f_{m+1} is the frequency of the group after the modal group

C is the group width or constant

$L_1 = 60.5$, $f_{bm-1} = 7$, $f_m = 8$, $f_{m+1} = 4$, C = 5

$$\text{Mode} = 60.5 + \left(\frac{8-7}{(8-7)+(8-4)}\right) \times 5$$

Mode = 60.5 + (1/5) × 5

Mode = 61.5

– •Summary

You have learned how to calculate the value of grouped mean, grouped mode and grouped median with given data was fully illustrated in this unit 2 and some examples were given for more understanding.

– •Self Assessment Questions

● You grew fifty baby carrots using special soil. You dig them up and measure their lengths (to the nearest mm) and group the results:

Length (mm)	Frequency
150 – 154	5
155 – 159	2
160 – 164	6
165 – 169	8
170 – 174	9
175 – 179	11
180 – 184	6
185 – 189	3

● The ages of the 112 people who live on a tropical island are grouped as follows:

Age	Number
0 -9	20
10 – 19	21
20 – 29	23
30 – 39	16
40 – 49	11
50 – 59	10
60 – 69	7
70 – 79	3
80 – 89	1

Obtain the mean, median and mode for grouped data

● Sammy caught ten rainbow trout, measured their lengths to the nearest inch, and recorded his results in groups as follows:

Length (in)	Number
15 – 19	2
20 – 24	7
25 – 29	1

Use the midpoints of the groups to estimate the mean length of the trout Sammy caught.

● Tommy trapped ten rabbits, weighed them to the nearest pound, and recorded his results in groups as follows:

Weight (lb)	Number
5 – 9	2
10 – 14	5
15 – 19	3

Use the midpoints of the groups to estimate the mean weight of the rabbits Tommy trapped.

● Thirty students in a class sat a science test. The results are recorded i groups as follows:

Mark	Number
20 - 29	1
30 – 39	1
40 – 49	10
50 – 59	11
60 – 69	5
70 - 79	2

Estimate the median mark correct to 1 decimal place.

Tutor Marked Assessment

- The numbers of words in each of the first eighty sentences of a book were counted.
  The results are recorded in groups as follows:

Number of Word	Number
1 – 4	2
5 – 8	5
9 – 12	11
13 - 16	23
17 - 20	21
21 – 24	13
25 – 28	4
29 – 32	1

Estimate the median length of sentence correct to 1 decimal place.

- The masses of 80 parcels were each measured to the nearest tenth of a kilogram, and the results recorded in groups as follows:

Mass (Kg)	Number
20.0 – 20.4	2
20.5 – 20.9	12
21.0 - 21.4	17
21.5 – 21.9	25
22.0 – 22.4	17
22.5 – 22.9	7

- Estimate the median mass correct to 1 decimal place.
  A frequency table displaying professor Blount's last statistic test is shown below:

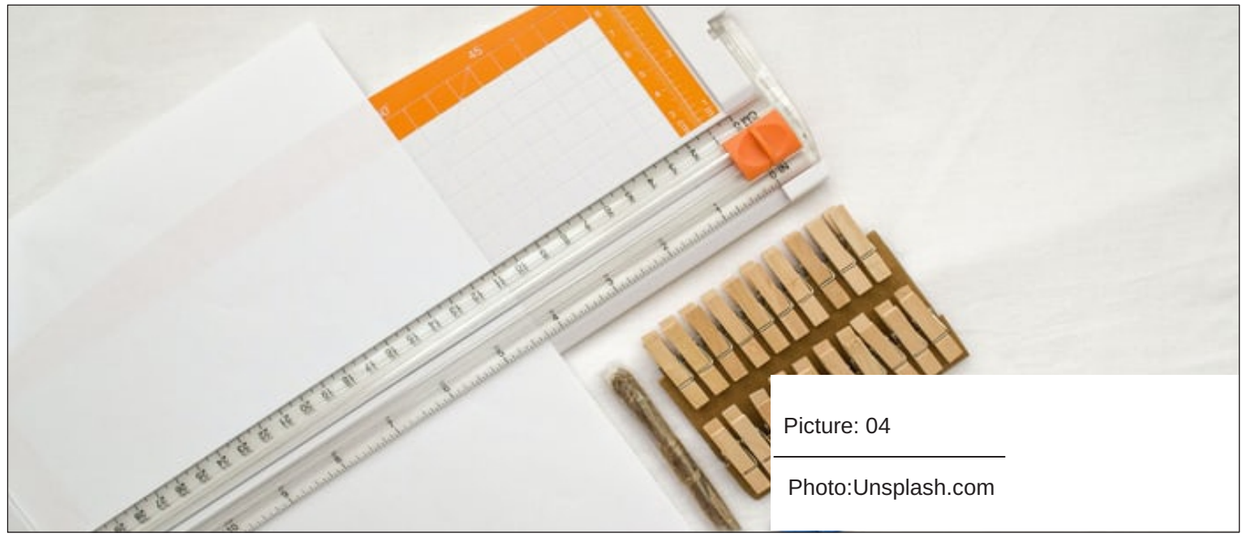
Grade Interval Number of Students

Grade Interval	Number of Students
50.0 – 56.5	1
56.5 – 62.5	0
62.5 – 68.5	4
68.5 – 74.5	4
74.5 – 80.5	2
80.5 – 86.5	3
86.5 – 92.5	4
92.5 – 98.5	1

Find the best estimate of the class mean.

– • References

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.
- Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.
- Gupta (2014). Introduction to Fundamental Statistics.
- Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.
- Introductory Statistics 3rd Edition by Sheldom M. Ross



Picture: 04

Photo:Unsplash.com

UNIT 3

Measures of dispersion or Measures variability

Introduction

The mean, median and most of the other "average" provide us with single numbers which represent whole sets of data. Other statistical measures we shall consider are called Measures of Variation. These measures, formally called statistics, are quavalues are determined by the data. We study the sample mean, sample median, and sample mode. These are all statistics that measure the center or middle value of a data set. Statistics that indicate the amount of variation in the data set are also considered. We learn about what it means for a data set to be normal, and we present an empirical rule concerning such sets.

Learning Outcomes

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

1. Relate the measures degree to which numerical values are spread around an average.
2. Discuss the following measures of dispersion Range, Mean deviation, variance, standard deviation, and semi – interquartile.
3. Calculate the coefficient of variation

Main Content

7 mins

You should be aware that data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data. In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called "Descriptive Statistics." You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs. A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data.

Range

Range is the difference between the largest data value and the smallest data value. For example, twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows:

5, 6, 3, 3, 2, 4, 8, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3.

Since the range is 6, we will keep each data value separate and not group them together. To create an ungrouped frequency distribution is a simple task. Place the data values from smallest to the largest without skipping any values on the first column. Place the **frequency**, the count of each data value, in the corresponding row of the second column.

Measures of Variation

It will interest you to know that an important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. In this section we will discuss three measures of variation: the range, the standard deviation and the variance.

Standard deviation

You should be aware that the most commonly used measure of variation, or spread, is the standard deviation. The standard deviation is a non-negative number that measures how far data values are from their mean. Its importance will become much clearer when we begin studying methods of Inferential Statistics. For now, we point out two key features of this important statistic.

The standard deviation provides a measure of the overall variation in a data set:

Standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation. Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes. Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average, and hence more predictable.

The standard deviation can be used to determine whether a data value is close to or far from the mean:

Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for
seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

The **variance is the average of the squares of the deviations.** For reasons that will become clear
later, we compute these averages slightly differently for population data than we do for sample data.

(2) The following data give the yearly numbers of law enforcement officers killed
in the United States over 10 years: 14, 15, 7, 14, 2, −3, −2, −19, −3, 5
Find the sample variance of the number killed in these years.

Solution

Rather than working directly with the given data, let us subtract the value 150
from each data item. (That is, we are adding c = −150 to each data value.)
This results in the new data set
Its sample mean is

$$\bar{Y} = \frac{14+15+7+14+2-3-2-19-3+5}{10}$$

$$= 3.0$$

$$S^2 = \frac{\sum_{i=1}^{n}(Y-\bar{Y})^2}{n-1}$$

$$S^2 = \sum Y^2 - n(\bar{Y})^2$$

$$= 1078 - 10(9)$$

$$= 988$$

Summary

In this unit:

- You have learnt how you can compute the measure of variation.

Self Assessment Questions

(1) Find the sample mean and the sample variance of this set 66, 68, 71, 72, 72, 75

(2)Find the sample mean and the sample variance of this set.
B: 2, 5, 9, 10, 10, 16

(3) An individual needing automobile insurance requested quotes from 10 different insurers for identical coverage and received the following values (amounts are annual premiums in dollars):
720, 880, 630, 590, 1140, 908, 677, 720, 1260, 800
Find

(a) The sample mean
(b) The sample median
(c) The sample standard deviation

(2) Compute the sample variance and sample standard deviation of the following data sets

(b) 6, 7, 8, 9, 10
Compute the sample variance and sample standard deviation of the following data sets:

(c) 11, 12, 13, 14, 15

Tutor Marked Assessment

- Find the sample standard deviation of the data set given by the following frequency table:

Value (A)	Frequency	Value (B)	Frequency
3	1	5	3
4	2	6	2

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X-\mu)^2}{N}$$ where, N = the number of data values in the population.

The sample variance is:

$$s^2 = \frac{\sum_{i=1}^{n}(X-\mu)^2}{n-1}$$

where n = the number of data values in the sample.

The symbol 2s represents the population variance and the population standard deviation σ is the square root of the population variance. Similarly, the symbol)(2*s* represents the sample variance, so the sample standard deviation (s) is the square root of the sample variance. Taking the square root of the variance can be thought of as reversing the effect of squaring the deviations. Thus, we can think of the standard deviation as a sort of "average" of the deviations. Note also that the standard deviation will always be measured in the same units as the data values.

SAQ 1

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X-\bar{X})^2}{N}}$$

The sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X-\bar{X})^2}{n-1}}$$

Illustrations

(1) In a fifth grade class, teacher was interested in the average age and the sample standard deviation of the
ages of her students. The following data are the ages for a sample of *n* = 20 fifth grade students.
The ages are rounded to the nearest half year:
9, 9.5, 9.5, 10, 10, 10, 10, 10.5, 10.5, 10.5, 10.5, 11, 11, 11, 11, 11, 11, 11.5, 11.5, 11.5.
The average age is 10.53 years, rounded to two places. Find the sample standard deviation.

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X-\bar{X})^2}{n-1}}$$

$$= \frac{(9-10.53)^2 + (9.5-10.53)^2 + \ldots + (11.5-10.53)^2}{20}$$

$$= \frac{9.7375}{19} = 0.5125$$

(2) Find the sample variance of data set 1, 2, 5, 6, 6.

$$S^2 = \frac{\sum_{i=1}^{n}(X-\mu)^2}{n-1}$$

$$= (1-4)^2 + \ldots + (6-4)^2$$

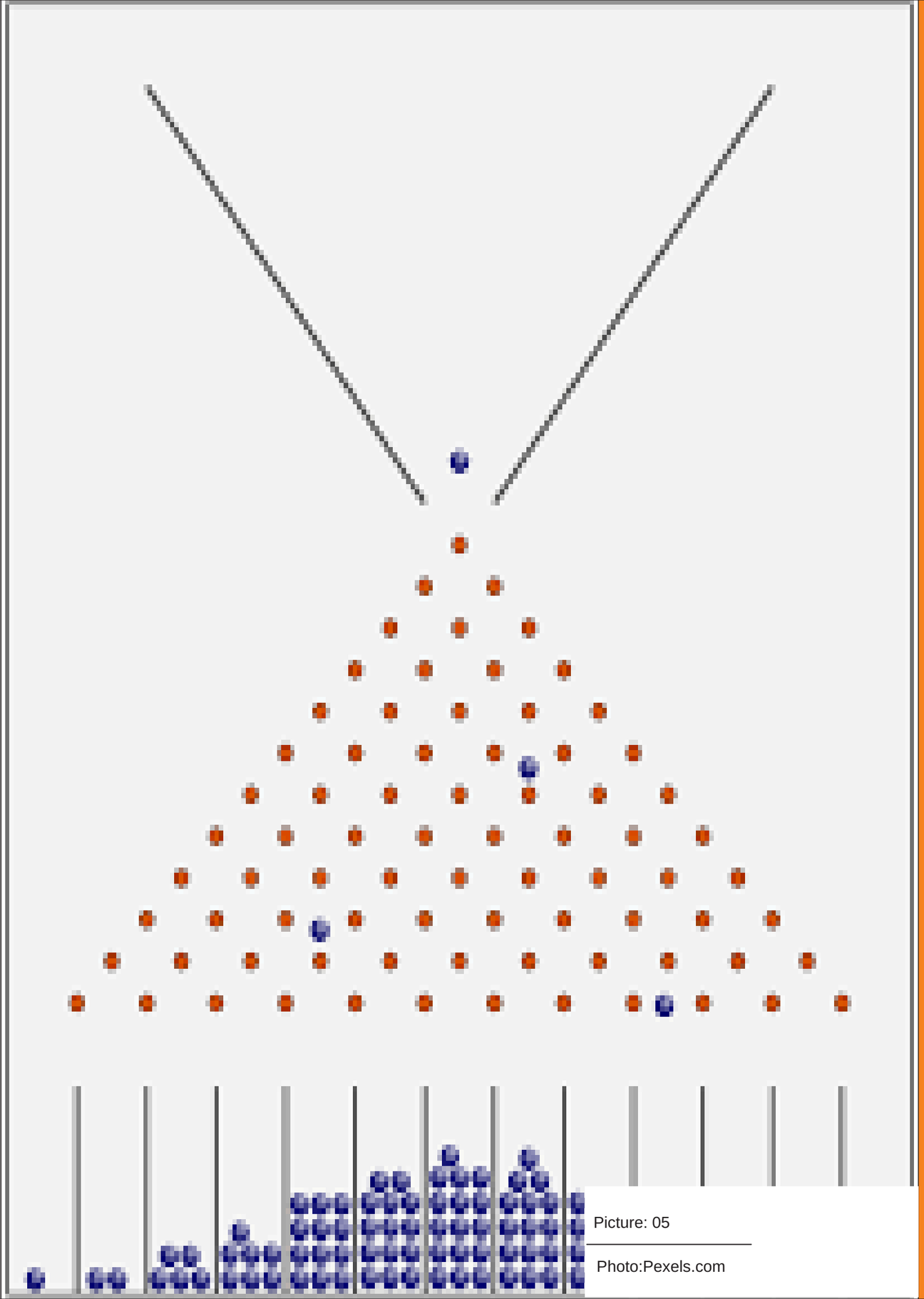
$$= 5.5$$

- Consider the following two data sets:
  A: 4.5, 0, 5.1, 5.0, 10, 5.2 B: 0.4, 0.1, 9, 0, 10, 9.5
- Determine the range for each data set.
- Determine the sample standard deviation for each data set.
- Determine the range for each data set.
- The following data represent the acidity of 40 successive rainfalls in the state of Minnesota. The acidity is measured on a pH scale, which varies from 1 (very acidic) to 7 (neutral).
  3.71, 4.23, 4.16, 2.98, 3.23, 4.67, 3.99, 5.04, 4.55, 3.24, 2.80, 3.44, 3.27, 2.66, 2.95, 4.70, 5.12, 3.77, 3.12, 2.38, 4.57, 3.88, 2.97, 3.70, 2.53,2.67, 4.12, 4.80, 3.55, 3.86, 2.51, 3.33, 3.85, 2.35, 3.12, 4.39, 5.09, 3.38, 2.73, 3.07

- Find the sample standard deviation.
- Find the range.
- The following data are the ages of a sample of 36 victims of violent crime in a large eastern city:
  25, 16, 14, 22, 17, 20, 15, 18, 33, 52, 70, 38, 18, 13, 22, 27, 19, 23, 33, 15, 13, 62, 21, 57, 66, 16, 24, 22, 31, 17, 20, 14, 26, 30, 18, 25
- Determine the sample standard deviation.

Reference

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and
- Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.
- Gupta (2014). Introduction to Fundamental Statistics.
- Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.
- Introductory Statistics 3rd Edition by Sheldom M. Ross

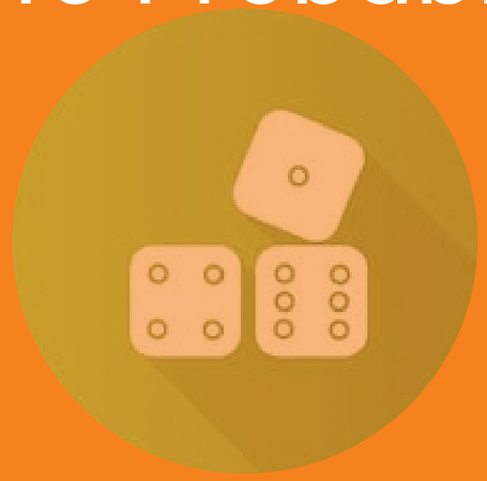


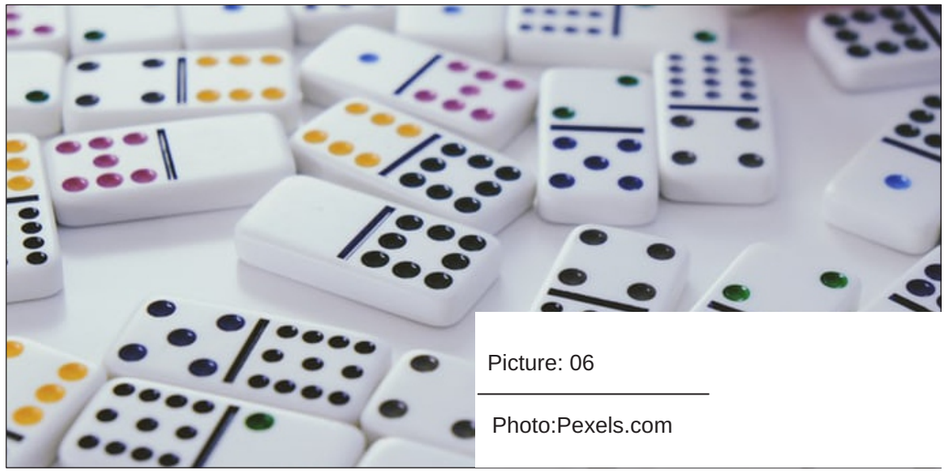
Picture: 05

Photo:Pexels.com

Module 2

Introduction To Probability





Picture: 06

Photo:Pexels.com

UNIT 1

Introduction to Probability

Introduction

In probability theory, we define a mathematical model of the above phenomenon by assigning “probabilities” (or the limit values or the relative frequencies) to the “events” connected with an experiment. Naturally, the reliability of our mathematical model for a given experiment depends upon the closeness of the assigned probabilities to the actual relative frequency.

Probability is the study of random or non-deterministic experiments. If a die is tossed in the air, then it is certain that the die will come down, but it is not certain that, say, a 6 will appear.

Learning Outcomes

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

1. Define probability.
2. Explain the terms of probability.
3. Express the axiom of probability.
4. Express mutually and non-mutually exclusive.

Main Content

7 mins

We will start with consideration of an experiment whose outcome cannot be predicted with certainty. We define the events of this experiment. We then introduce the concept of the probability of an event, which is the probability that the outcome of the experiment is contained in the event. An interpretation of the probability of an event as being a long-term relative frequency is given. Properties of probabilities are discussed. The conditional probability of one event, given the occurrence of a second event, is introduced. We see what it means for events to be independent.

The word probability is a commonly used term that relates to the chance that a particular event will occur when some experiment is performed, where we use the word experiment in a very broad sense. Indeed, an experiment for us is any process that produces an observation or outcome. We are often concerned with an experiment whose outcome is not predictable, with certainty, in advance. Even though the outcome of the experiment will not be known in advance, we will suppose that the set of all possible outcomes is known. This set of all possible outcomes of the experiment is called the sample space and is denoted by S.

We say that *A* and *B* are mutually exclusive events if they cannot occur at the same time. This means that *A* and *B* do not share any outcomes and so it follows that P(*A* and *B*) = 0. The non- mutually exclusive events means that event A and event B share outcome or event in common.

Three properties

(1) Property 1: For any event A, the probability of A is a number between 0 and 1. That is, $0 \leq P(A) \leq 1$

(2) Property 2: The probability of sample space S is 1. Symbolically, P(S) = 1

(3) Property 3: The probability of the union of disjoint events is equal to the sum of the probabilities of these events. For instance, if A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$

Illustration

Let's look at some examples.

Example 1

In a survey of 420 members of a retirement center, it was found that 144 are smokers and 276 are not. If a member is selected in such a way that each of the members is equally likely to be the one selected, what is the probability that person is a smoker?

Solution

Note that there are 420 outcomes in the sample space of the experiment of selecting a member of the center. Namely, the outcome is the person selected. Since there are 144 outcomes in the event that the selected person is a smoker, it follows that the probability of this event is $P\{\text{smoker}\} = \frac{144}{420} = \frac{12}{35}$

Example 2

Suppose that when two dice are rolled, each of the 36 possible outcomes is equally likely. Find the probability that the sum of the dice is 6 and the sum is 7.

Solution

If we let A denote the event that the sum of the dice is 6 and B that it is 7, then

A = {(1, 5),(2, 4),(3, 3),(4, 2),(5, 1)}

and

B = {(1, 6),(2, 5),(3, 4),(4, 3),(5, 2),(6, 1)}

Therefore, since A contains 5 outcomes and B contains 6, we see that

P(A) = P{sum is 6} = 5/36

P(B) = P{sum is 7} = 6/36 = 1/6

Example 3

An elementary school is offering two optional language classes, one in French and the other in Spanish. These classes are open to any of the 120 upper-grade students in the school. Suppose there are 32 students in the French class, 36 in the Spanish class, and a total of 8 who are in both classes. If an upper-grade student is randomly chosen, what is the probability that this student is enrolled in at least one of these classes?

Solution

Let A and B denote, respectively, the events that the randomly chosen student is enrolled in the French class and is enrolled in the Spanish class. We will determine P(A ∪ B), the probability that the student is enrolled in either French or Spanish, by using the addition rule

P(AUB)=P(A)+P(B)-P(A∩B)

Since 32 of the 120 students are enrolled in the French class, 36 of the 120 are in the Spanish class, and 8 of the 120 are in both classes, we have

$P(A) = \frac{32}{120}$ $P(B) = \frac{36}{120}$ $P(A \cap B) = \frac{8}{120}$

Therefore,

$$P(A \cup B) = \frac{32}{120} + \frac{36}{120} - \frac{8}{120} = \frac{1}{2}$$

That is, the probability that a randomly chosen student is taking at least one of the language classes is 1/2.

Example4

One man and one woman are to be selected from a group that consists of 10 married couples. If all possible selections are equally likely, what is the probability that the woman and man selected are married to each other?

Solution

Once the man is selected, there are 10 possible choices of thewoma since one of these 10 choices is the wife of the man chosen, we see that desired probability is 1/ 10

When each outcome of the sample space is equally likely to be the outcome of the experiment, we say that an element of the sample space is randomly selected.

(4) The lists the earnings frequencies of all full-time workers who are at least 15 years old, classified according to their annual salary and gender.

Earnings of Workers by Sex, 1989

Earnings group	Number		Distribution (percent)	
(in $1000)	Women	Men	Women	Men
<5	427,000	548,000	1.4	1.1
5–10	440,000	358,000	1.4	0 .7
10–15	1,274,000	889,000	4.1	1.8
15–20	1,982,000	1,454,000	6.3	2.9
20–30	6,291,000	5,081,000	20.1	10.2
30–40	6,555,000	6,386,000	20.9	12.9
40–50	5,169,000	6,648,000	16.5	13.4
50–100	8,255,000	20,984,000	26.3	42.1
>100	947,000	7,377,000	3.0	14.9
Total	31,340,000	49,678,000	100.0	100.0

Source: Department of Commerce, Bureau of the Census.

Solution

Suppose one of these workers is randomly chosen. Find the probability that this

person is.....

(a) A woman (b) A man (c) A man earning under $30,000 (d) A woman earning over $50,000

Solution

(a) Since 31,340,000 of the 31,340,000 + 49,678,000 = 81,018,000 workers are women, it follows that the probability that a randomly chosen worker is a woman is

3868.081018000000,340,31=

That is, there is approximately a 38.7 percent chance that the randomly selected worker is a woman.

(b) Since the event that the randomly selected worker is a man is the complement of the event that the worker is a woman, we see from (a) that the probability is approximately 1 − 0.3868 = 0.6132.

(c) Since (in thousands) the number of men earning under $30,000 is

548 + 358 + 889 + 1454 + 5081 = 8330

we see that the desired probability is 8330/81,018 ≈ .1028. That is, there is approximately a 10.3 percent chance that the person selected is a man with an income under $30,000.

(d) The probability that the person selected is a woman with an income above $50,000 is

1136.0810189478255=+

That is, there is approximately an 11.4 percent chance that the person selected is a woman with an income above $50,000.

(1) As a further check of the preceding formula for the conditional probability, use it to compute the conditional probability that the sum of a pair of rolled dice is 10, given that the first die lands on 4.

Solution

Letting B denote the event that the sum of the dice is 10 and A the event that the first die lands on 4, we have

$$P(B|A) = \frac{P(A \cap B)}{P(A)})$$

$$= \frac{P(\{(4,6)\}}{P(\{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\})}$$

$$= \frac{1/36}{6/36}$$

$$= \frac{1}{6}$$

The organization that employ jacobi is organizing a parent daughter dinner for those employee at least having 1 daughter. each of this employee to have two children, what is the conditional probability that those both are girls given by Jacobi is invited for the dinner. Assume the sample space S is given by

$$S = \{(g,g),(g,b),(b,g),(b,b)\}$$

and that all these outcomes are equally likely, where the outcome (g, b) means s,for instance, that Jacobi's oldest child is a girl and youngest is a boy.

Solution

Since Jacobi is invited to the dinner, we know that at least one of Jacobi's children is a girl. Letting B denote the event that both of them are girls and A the event that at least one is a girl, we see that the desired probability is P(B|A). This is determined as follows:

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{P(g,g)}{P(\{(g,g),(g,b),(b,g)\})}$$

$$= \frac{1/4}{3/4} = \frac{1}{3}$$

(7) Suppose the number (in thousands) of students enrolled in a California State College, categorized by sex and age.

(a) Suppose a student is randomly chosen. What is the probability this student is a women?

Find the conditional probability that a randomly chosen student is

(b) Over 35, given that this student is a man

(c) Over 35, given that this student is a women

(d) A women, given that this student is over 35

(e) A man, given that this student is between 20 and 21.

Solution

(a) Since there are 6663 women out of a total of 12,544 students, it follows that the probability that a randomly chosen student is a women is $\frac{6663}{12,544}$

(b) Since there are a total of 5881 males, of whom 684 are over age 35, the desired conditional probability is P(over 35/man) = $\frac{684}{5881} = 0.1163$

(c) By similar reasoning to that used in (b) we see tha P(over 35/women) = $\frac{1339}{6663} = 0.2010$

(d) Since there are a total of 684 + 1339 = 2023 students who are over age 35, of whom 1339 are women, it follows that P(women/over 35) = $\frac{1339}{2023} = 0.6619$

(e) Since there are a total of 1089 + 1135 = 2224 students who are between 20 and 21 of whom 1089 are men, it follows that P(man/between 20 and 21) = $\frac{1089}{2224} = 0.4897$

Since $P(B/A) = \frac{P(A \cap B)}{P(A)}$

we obtain, upon multiplying both sides by P(A), the following result, known as the multiplication rule.

Summary

- We try to show how probability was determine and obtained through the event that happen.

Self Assessment Questions

(1) Define probability.

(2) Explain the terms of probability.

(3) Express the axiom of probability.

(4) Express mutually and non-mutually exclusive.

Tutor Marked Assessment

- A student is chosen at random to represent a class with five freshmen, eight sophomores, three juniors, and two seniors. Find the probability that the student is (a) sophomores, (b) a junior, (c) a junction or a senior

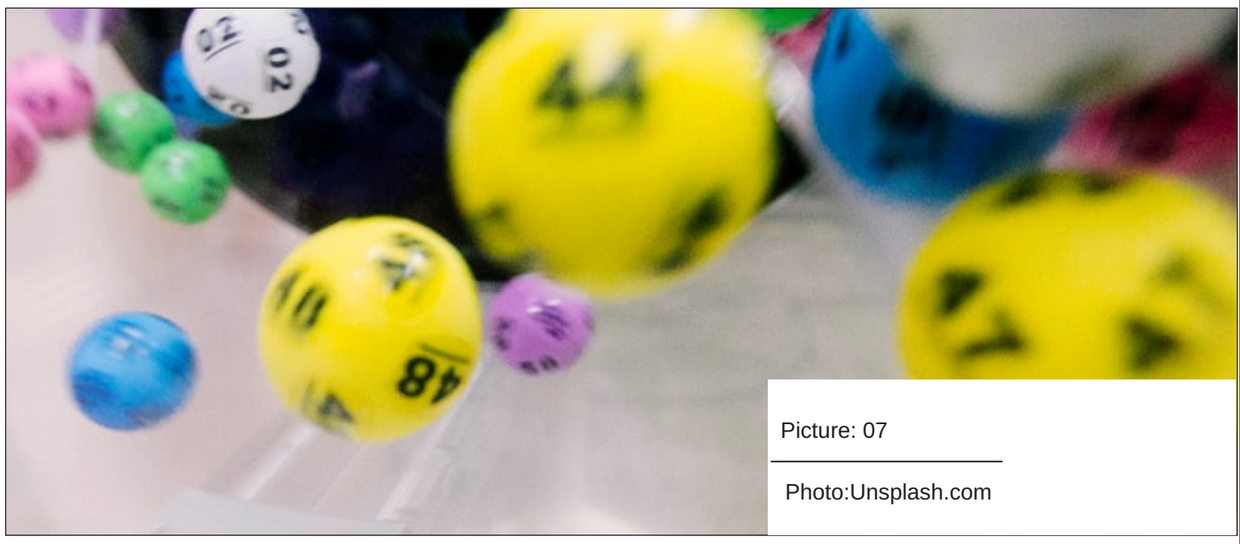
- Of 10 girls in a class, 3 have blue eyes. Two of the girls are chosen at random. Find the probability that (a)both have blue eyes, (b) neither has blue eyes, (c) at least one has blue eyes, (d) exactly one has blue eyes.

- Three bolts and three nuts in a box. Two are chosen at random. Find the probability that one is a bolt and one is a nut.

- A box contains two white sox, two blue sox, and two red sox. Two sox are drawn at random. Find the probability they are a match (same color).
Of 120 students, 60 are studying French, 50 are studying Spanish, and 20 are studying both French and Spanish. A student is chosen at random.
- Find the probability that the student is studying: (a) French or Spanish, (b) neither French nor Spanish, (c) only French, (d) exactly one of the two languages.

References

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.
- Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.
- Gupta (2014). Introduction to Fundamental Statistics.
- Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.
- Introductory Statistics 3rd Edition by Sheldom M. Ross



Picture: 07

Photo:Unsplash.com

UNIT 2

Experimental Probability

Introduction

The mathematical theory of probability for finite sample spaces provides a set of number called weights, ranging 0 to 1. To every point in the sample space be assign a weight such that the sum of all the weight is 1. To find the probability of any event A we sum all weights assigned to the sample point in A. This sum is called the measure of A or the probability of A and is defined by P(A). Independent and dependent event. Therefore the following are the axioms of probability: (i) 01)(££*AP* (ii) 0)(=f*P* (iii) 1)(=*SP*

Learning Outcomes

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

1. Obtain experimental probability.
2. Compute the tossing of coin or a die.
3. Calculate the Independent and dependent event.

Main Content

8 mins

It will interest you to know that probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An experiment is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a chance experiment. Flipping a fair coin twice is an example of an experiment. A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all
possible outcomes. We will discuss three ways to represent a sample space are: list the possible
outcomes, create a tree diagram, or using a Venn diagram. The uppercase letter *S* is used to denote the sample space. For example, if you flip one fair coin, $S = \{H, T\}$ where *H* = heads and
T = tails are the outcomes. Sample spaces can often be written in multiple ways. For example, if
We toss a coin three times in succession, we could describe the sample space in terms of the
Number of heads as $S = \{0, 1, 2, 3\}$ or we can list the possible sequences of heads and tails as $S = \{HHH, THH, HTH, HHT, HTT, THT, TTH, TTT\}$.
We will see in a moment that the second way is more useful, since it has **equally likely outcomes.** That is, each of the eight outcomes is equally likely to occur. It should be intuitively clear that this is not true of the first set; i.e. there is only one way for all three tosses to land as heads, but there is more than one way to get 1 head in three tosses.
An **event** is any collection of outcomes. Upper case letters like *A* and *B* represent events. For
example, if the experiment is to flip one fair coin, event *A* might be getting at most one head.
The probability of an event A is written as P(A).
There are two basic ways to calculate probabilities; Theoretical Probability and Empirical Probability.

Theoretical Probability

Suppose that all outcomes in the sample space are equally likely. To calculate the probability of an event A, count the number of outcomes for event A and divide by
The total number of outcomes in the sample space. That is,

$$P(A = \frac{number\ \ of\ \ outcome\ \ inA}{number\ \ of\ \ outcome\ \ inS})$$

Empirical Probability

The probability of any event is the **long-term relative frequency** of that event. That is, if we were to do a very large number of trials (repetitions of our experiment) then the probability of A is

$$P(A) = \frac{number\ \ of\ \ timesA\ \ occurs}{total\ \ number\ \ of\ \ trials}$$

For example, suppose that an insurance company wants to find the probability that a suburban male driver, whose age is between 20 and 25 years, will have an accident in the coming year.
Then they would use data from thousands of such drivers and calculate the proportion of those drivers that had an accident in the past year.

Properties of Probability

The following are some of the properties of probability.

(1) For any event *A*, $\mathbf{0 < P(A) < 1}$. That is, the probability of any event is always a number between o and 1

(2) The probability of any event *A* is equal to the sum of the probabilities of the individual Outcomes in A

(3) The sum of the probabilities of all outcomes in *S* must equal 1. This is true whether or not *S* consists of equally likely outcomes.

Illustration

Some examples are given below:

Example 1

Suppose that we roll two fair dice, one red and one white. Then there are 36 possible outcomes in the sample space. Suppose we write the sample space in terms of the number of dots facing up on the two dice.

That is, write S = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}. Use the table to answer each of the following:

(a) Are the outcomes in S equally likely?

(b) Find P(4).

(c) Find P(6)

(d) Let E be the event "number of dot is even". Find P(E)

(e) Let G be the event "number of dot is greater than 9" Find P(G).

(f) Suppose that we rolled the dice 7200 times. How many times would we expect to roll a 7?

Solution

The events are not equally likely – for example the probability of rolling "snake eyes" is

(a) P(2) = 1/36 whereas the probability of rolling a 3 is P(3) = 2/36.

(b) There are 3 outcomes in which the dots add to 4 so P(4) = 3/36.

(c) There are 5 outcomes in which the dots add to 6 so P(6) = 5/36.

(d) The event E = {2, 4, 6, 8, 10, 12}. So we add the probabilities of the outcomes

$$P(E) = P(2) + P(4) + P(6) + P(8) + P(10) + P(12)$$

$$= \frac{1}{36} + \frac{3}{36} + \frac{5}{36} + \frac{5}{36} + \frac{3}{36} + \frac{1}{36} = \frac{18}{36}$$

(e) The event G = {10, 11, 12}. So we add the prob $\frac{6}{36} = \frac{1}{6}$ of the outcomes:

$$P(E) = P(2) + P(4) + P(6) + \frac{6}{36} = \frac{1}{6} + P(10) + P(12)$$

$$= \frac{1}{36} + \frac{3}{36} + \frac{5}{36} + \frac{5}{36} + \frac{3}{36} + \frac{1}{36} = \frac{18}{36}$$

(f) The theoretical probability of rolling a 7 is $P(7) = \frac{6}{36} = \frac{1}{6}$

Example 2

The table below describes the distribution of a random sample S of 100 individuals, organized by

gender and whether they are right – or left – handed.

	Right - handed	Left – handed
Males	43	9
Females	44	4

Let's denote the events as M = the subject is male, F = the subject is female, R = the subject is right – handed and L = the subject is left – handed. Compute the following probabilities.

(a) P(M) (b) P(F) (c) P (R) (d) P(L) (e) P(M and R) (f) P(F and L) (g) P(M or F) (h) P(M or R) (i) P(F or L) (j) P (M')

Solution

(a) $P(M) = 0.52$ (b) $P(F) = 0.48$ (c) $P(R) = 0.87$ (d) $P(L) = 0.13$

(e) $P(M \text{ and } R) = 0.43$ (f) $P(F \text{ and } L) = 0.04$ (g) $P(M \text{ or } F) = 1$

(h) P(M or R) = $\frac{43+9+44}{100} = 0.960$

(I) P (F or L) = $\frac{44+4+9}{100} 0.57$

(j) P(M') = P(F) = 0.48

Example 3

(3) In a large metropolitan area, it is known that 45% of all voters are registered as Democrats.

Suppose we select two voters at random.

a. What is the probability that both are registered as Democrats?

b. What is the probability that neither are registered as Democrats?

Solution

Let event $D1$ = Democrat is selected on the first choice, and $D2$ = Democrat is selected on the second**.** The key observation here is that the draws are independent, so we can use the rule

$P(D1 \text{ and } D2) = P(D1)P(D2)$.

a. The probability that both are registered as Democrats is:

$P(D1 \text{ and } D2) = P(D1)\ P(D2) = 0.45 \times 0.45 = 0.2025$.

b. If 45% are registered as Democrats, then the remaining 55% are *not* registered as Democrats.

$P(D1' \text{ and } D2') = P(D1')\ P(D2') = 0.55 \times 0.55 = 0.3025$.

Example 4

Let G = event that a student is taking a math class. Let H = event that a student is taking a science class. Then, G and H = the event that a student is taking *both* a math class and a science class.

Suppose that P(G) = 0.6, P(H) = 0.5, and P(G and H) = 0.3. Based on this information, are G and H independent?

Solution

To show that G and H are independent, we must show **ONE** of the following:

P(G | H) = P(G)

P(H | G) = P(H)

P(G and H) = P(G)P(H)

The option we choose depends on the information given in the problem. We could choose any of the methods here because we have the necessary information.

For example, we can show that P(G | H) = P(G):

$$P(G/H) = \frac{P(G \cap H)}{P(H)} = \frac{0.3}{0.5} = 0.6 = P(G)$$

Or, we can show P(G and H) = P(G)P(H):

P(G)P(H) = (0.6)(0.5) = 0.3 = P(G and H)

Since G and H are independent, knowing that a person is taking a science class does not affect the probability that he or she is taking a math class.

Example 5

In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. One student is picked randomly. Let F be the event that a student is female, and let L be the event that a student has long hair. Are the events F and L independent?

There are three conditions we can check to determine independence:

P(F) = P(F | L)

or P(F and L) = P(F)P(L)

The one we use will depend on the information given in the problem. So we first write down the probabilities that are given in the problem:

P(F) = 0.60; P(L) = 0.50; P(F AND L) = 0.45; P(L | F) = 0.75.

Based on this information we could use either the second condition or the third.

(We do not know $P(F \mid L)$ yet, so we cannot use the first condition.)

Using the second rule, we check whether P(L | F) equals P(L):

We are given that P(L |F) = 0.75, whereas P(L) = 0.50. Since these are not equal, the events are *not* independent. This shows that a female student is more likely to have long hair than is a male student. Using the third rule, we check whether P(F and L) = P(F)P(L):

We are given that P(F and L) = 0.45, whereas P(F)P(L) = (0.60)(0.50) = 0.30. Since these are not equal, the events F and L are not independent.

(5) Suppose that we toss two fair coins. Find the probabilities of the events.

(a) Let F = the event of getting at most one tail (zero or one tail).

(b) Let G = the event of getting two faces that are the same.

(c) Let H = the event of getting a head on the first flip followed by a head or tail on the second flip.

(d) Are F and G mutually exclusive?

(e) Let J = the event of getting all tails. Are J and H mutually exclusive?

Solution

(a) Zero (0) or one (1) tails occur when the outcomes HH, TH, HT show up. P(F)= 3/4.

(b) Two faces are the same if HH or TT show up. P(G) = 2/4 = 1/2.

(c) A head on the first flip followed by a head or tail on the second flip occurs when HH or HT show up. So P(H) = 2/4 = 1/2.

(d) F and G share the outcome HH so F and G are not mutually exclusive.

(e) Getting all tails occurs when tails shows up on both coins (TT). H's outcomes are HH and HT. J and H have no outcomes in common, J and H are mutually exclusive.

Summary

- In this unit, you have learnt:
- How to obtained experimental probability
- The tossing of coin or a die and
- How to calculate the Independent and dependent event. .

Self Assessment Questions

(1) Obtain experimental probability.

(2) Compute the tossing of coin or a die.

(3) Calculate the Independent and dependent event.

Tutor Marked Assessment

- A lot contains 12 items of which 4 are defective. Three items are drawn at random from the lot one after the other without replacement. Find the probability that all the three are non – defective.
- A pair of dice is rolled. If the sum on the two dice is 9, find the probability that one of the dice showed 3.
- In a single throw of two dice, what is the probability of getting a total different from 8.
- The probability that a patient recovers from a rare blood disease is 0.4. If 15 people are known to have contracted this disease, what is the probability that at least 10 survive.
- The probability that a patient recovers from a rare blood disease is 0.4. If 15 people are known to have contracted this disease, what is the probability that exactly 5 survive.
- An electronic company manufactures a specific component for an ultrasound machine. It finds that out of every 1000 components produced, 8 are defective. If the components are packed in batches of 250 find the probability of obtaining exactly two defective in a batch.

The number of deaths by road, as recorded in a company, is 2 per 50,000 of the population. Find the probability in a town of 100,000 in habitants within the country exactly five deaths by road, will be recorded.

- The number of deaths by road, as recorded in a company, is 2 per 50,000 of the population. Find the probability in a town of 100,000 in habitants within the country between one and three deaths by road, inclusive will be recorded.

- If a fair coin is tossed six times, find the probability of obtaining at least two heads.
- A fair die and a coin are rolled at once, find the probability of getting a tail and an even number

References

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.
- Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.
- Gupta (2014). Introduction to Fundamental Statistics.
- Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.
- Introductory Statistics 3rd Edition by Sheldom M. Ross.

Picture: 08

Photo:Unsplash.com

UNIT 3

Law of Probability

Introduction

There are basically two laws of probability namely

(1) Additional law of probability

(2) Multiplication law of probability

Addition law of probability is used when we want to calculate the probability of one event or the other. For two events E_1 or E_2 is symbolically written as follows:

$P(E_1 \text{ or } E_2) = P(E_1 \cup E_2)$.

Multiplication law of probability states that the probability of a combined occurrence of two (or more) events E_1 and E_2 is the product of the probability E_1 and the conditional probability of E_2 on the assumption that E_1 has occurred. This is denoted by $P(E_1 \text{ and } E_2)$ or simply, $P(E_1 \cap E_2) = P(E_1).P(E_2/E_1)$.

Where $P(E_2/E_1)$ is the conditional probability of event E_2 on the assumption that E_1 occurs at the same time.

Learning Outcomes

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

1. Know the two laws of probability.
2. Relate mutually exclusive event.
3. Detect independent and dependent event.

Main Content

7 mins

You should not that if we interpret P(A) as the long-run relative frequency of event A, then the stated conditions are satisfied. The proportion of experiments in which the outcome is contained in A would certainly be a number between 0 and 1. The proportion of experiments in which the outcome is contained in S is 1 since all outcomes are contained in sample space S. Finally, if A and B have no outcomes in common, then the proportion of experiments whose outcome is in either A or B is equal to the proportion whose outcome is in A plus the proportion whose outcome is in B. For instance, if the proportion of time that a pair of rolled dice sums to 7 is 1/6 and the proportion of time that they sum to 11 is 1/18, then the proportion of time that they sum to either 7 or 11 is 1/6 + 1/18 = 2/9.

In words, the probability that the outcome of the experiment is not contained in A is 1 minus the probability that it is. For instance, if the probability of obtaining heads on the toss of a coin is 0.4, then the probability of obtaining tails is 0.6.

The following formula relates the probability of the union of events A and B, which are not necessarily disjoint, to P(A), P(B), and the probability of the intersection of A and B. It is often called the addition rule of probability. To see why the addition rule holds, note that P(A ∪ B) is the probability of all outcomes that are either in A or in B. On the other hand, P(A) + P(B) is the probability of all the outcomes that are in A plus the probability of all the outcomes that are in B. Since any outcome that is in both A and B is counted twice in P(A) +P(B).

Illustration

Example 1

Suppose that when two dice are rolled, each of the 36 possible outcomes in is equally likely. Find the probability that the sum of the dice is 6 and that it is 7.

Solution

If we let A denote the event that the sum of the dice is 6 and B that it is 7, then
A = {(1, 5),(2, 4),(3, 3),(4, 2),(5, 1)}and B = {(1, 6),(2, 5),(3, 4),(4, 3),(5, 2),(6, 1)}

Therefore, since A contains 5 outcomes and B contains 6, we see that
P(A) = P{sum is 6} = 5/36

P(B) = P{sum is 7} = 6/36 = 1/6

Example 2

An elementary school is offering two optional language classes, one in French and the other in Spanish. These classes are open to any of the 120 upper-grade students in the school. Suppose there are 32 students in the French class, 36 in the Spanish class, and a total of 8 who are in both classes. If an upper-grade student is randomly chosen, what is the probability that this student is enrolled in at least one of these classes?

Solution

Let A and B denote, respectively, the events that the randomly chosen student is enrolled in the French class and is enrolled in the Spanish class. We will determine P(A ∪ B), the probability that the student is enrolled in either French or Spanish, by using the addition rule P(A ∪ B) = P(A) + P(B) − P(A ∩ B)

Since 32 of the 120 students are enrolled in the French class, 36 of the 120 are in the Spanish class, and 8 of the 120 are in both classes, we have

$$P(A)=\frac{32}{120} \quad , \quad P(B)=\frac{36}{120} \quad , P(A\cap B)=\frac{8}{120}$$

Therefore,

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

$$P(A\cup B)=\frac{32}{120}+\frac{36}{120}-\frac{8}{120}=\frac{1}{2}$$

That is, the probability that a randomly chosen student is taking at least one of the language classes is 1/2.

Example 3

One man and one woman are to be selected from a group that consists of 10 married couples. If all possible selections are equally likely, what is the probability that the woman and man selected are married to each other?

Solution

Once the man is selected, there are 10 possible choices of the woman. Since one of these 10 choices is the wife of the man chosen, we see that the desired probability is 1/10. when each outcome of the sample space is equally likely to be the outcome of the experiment, we say that an element of the sample space is randomly selected.

Example 4

In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. One student is picked randomly. Let F be the event that a student is female and let L be the event that a student has long hair. Are the events F and L independent?

Solution

There are three conditions we can check to determine independence:

$P(F) = P(F \mid L)$

$P(L) = P(L \mid F)$

or $P(F \text{ and } L) = P(F)P(L)$

The one we use will depend on the information given in the problem. So we first write down the probabilities that are given in the problem: $P(F) = 0.60$; $P(L) = 0.50$; $P(F \text{ AND } L) = 0.45$; $P(L \mid F) = 0.75$.

Based on this information we could use either the second condition or the third.

(We do not know $P(F \mid L)$ yet, so we cannot use the first condition.)

Using the second rule, we check whether $P(L \mid F)$ equals $P(L)$:

We are given that $P(L \mid F) = 0.75$, whereas $P(L) = 0.50$. Since these are not equal, the events are *not* independent. This shows that a female student is more likely to have long hair than is a male student.

Using the third rule, we check whether $P(F \text{ and } L) = P(F)P(L)$:

We are given that $P(F \text{ and } L) = 0.45$, whereas $P(F)P(L) = (0.60)(0.50) = 0.30$. Since these are not equal, the events F and L are not independent.

(5) Let event C = taking an English class. Let event D = taking a speech class. Suppose $P(C) = 0.75$,

$P(D) = 0.3$, $P(C \mid D) = 0.75$ and $P(C \text{ and } D) = 0.225$. Use this information to answer the following:

a. Are C and D independent?
b. Are C and D mutually exclusive?
c. What is $P(D \mid C)$?

Solution

(a) Yes, because $P(C \mid D) = P(C)$.

(b) No, because $P(C \text{ and } D) \neq 0$.

(c) $P(D/C) = \frac{P(C \cap D)}{P(C)} = \frac{0.225}{0.3} = 0.75$

Example 6

Klaus is trying to choose where to go on vacation. Klaus can only afford one vacation.

His two choices are: A = New Zealand and B = Alaska. The probability that he chooses A is

$P(A) = 0.6$ and the probability that he chooses B is $P(B) = 0.35$. $P(A \text{ and } B) = 0$, because Klaus

can only afford to take one vacation. Therefore, the probability that he chooses either New Zealand or Alaska is:

Solution

$$P(AorB) = P(A) + P(B) = 0.6 + 0.35 - 0 = 0.95$$

Example 7

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two in a row in the next game. Let A = the event Carlos is successful on his first attempt, so:

P(A) = 0.65. Let B = the event Carlos is successful on his second attempt. P(B) = 0.65.

Carlos tends to shoot in streaks; the probability that he makes the second goal *given* that he made the first goal is 0.90.

a. What is the probability that he makes both goals?

b. What is the probability that Carlos makes either the first goal or the second goal?

c. Are A and B independent?

d. Are A and B mutually exclusive?

(a) The problem is asking you to find $P(AA \cap B) = P(BA \cap A)$. Since P($B|A$) = 0.90:

$P(B \cap A) = P(B/A)P(A)$

$P(B \cap A) = P(B/A)P(A)$

$=(0.90)(0.65)$

$= 0.585$

Carlos makes the first and second goals with probability 0.585.

(b) The problem is asking you to find P(A OR B).

$P(AORB) = P(A) + P(B) - P(A \cap B)$

$=0.65+0.65-0.585$

$= 0.715$

Carlos makes either the first goal or the second goal with probability 0.715.

(c) No, they are not independent events, because $P(B \cap A) = 0.585$.

P(B)·P(A) = (0.65)(0.65) = 0.423

$0.423 \neq 0.585 = P(B \cap A)$ So, $P(B \cap A)$ is not equal to P(B)·P(A)

(d) No, they are not mutually exclusive because P(A and B) = 0.585.
To be mutually exclusive P(A and B) = 0.

Example 9

A community swim team has 150 members. Seventy-five of the members are advanced swimmers. Forty- seven of the members are intermediate swimmers. The rest are novice swimmers. Forty of the advanced swimmers practice four times a week. Thirty of the intermediate Swimmers practice four times a week. Ten of the novice swimmers practice four times a Week.

Suppose one member of the swim team is chosen randomly.

(a) What is the probability that the member is a novice swimmer?

(b) What is the probability that the member practices four times a week?

(c) What is the probability that the member is an advanced swimmer and practices four times a week

(d) What is the probability that a member is an advanced swimmer and an Intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?

(e) Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Solution

(a) 28/150

(b) 80/150

(c) 40/150

(d) P(advanced and intermediate) = 0, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

(e) No, these are not independent events.
P(novice and practices four times per week) = 0.0667
P(novice)P(practices four times per week) = 0.0996
Since 0.0667 ≠ 0.0996, the events are not independent.

Summary

In this unit, you have learnt:

- The two laws of probability,
- Mutually exclusive event and
- Independent and dependent event.

Self Assessment Questions

(1) Know the two laws of probability.

(2) Relate mutually exclusive event.

(3) Detect independent and dependent event.

Tutor Marked Assessment

- Felicity attends Modesto JC in Modesto, CA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class given that she enrolls in speech class is 0.25. Let: M = math class, S = speech class.
- Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time. Let B = woman develops breast cancer and let N = tests Negative. Suppose one woman is selected at random.
- What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?
- Given that the woman has breast cancer, what is the probability that she tests negative?
- What is the probability that the woman has breast cancer and tests negative?
- What is the probability that the woman has breast cancer or tests negative?
- Are having breast cancer and testing negative independent events? Are having breast cancer and testing negative mutually exclusive?

- Forty percent of the students at a local college belong to a club and **50%** work part time. Five percent of the students work part time and belong to a club. Draw a Venn diagram displaying this information, letting C = "student belongs to a club" and PT = "student works part time".

If a student is selected at random, find:

- The probability that the student belongs to a club **given** that the student works part time.
- The probability that the student belongs to a club **or** works part time
- The probability that a student works part time, but does not belong to a club.
- The probability that a student neither belongs to a club nor works part time.
- A person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. It is known that 4% of African Americans have both type O blood and a negative Rh factor, 8% of African Americans have the Rh- factor, and 51% type O blood. Make a Venn diagram for this situation, using O for the set of individuals with type O blood, and Rh- for the individuals with the negative Rh factor.
- Find the probability that a randomly selected African American has type O blood or negative Rh factor.
- Find the probability that a randomly selected African American has negative Rh factor, but does not have type O blood.
- Find the probability that a randomly selected African American has neither type O blood nor a negative Rh factor.

- A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder is taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the Seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?
- Given that a woman develops breast cancer, what is the probability that she tests positive? Find $P(P|B) = 1 - P(N|B)$.

References

Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.

Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.

Gupta (2014). Introduction to Fundamental Statistics.

Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.

Introductory Statistics 3rd Edition by Sheldom M. Ross.

Picture: 09

Photo:Unsplash.com

Module 3

Binomial Distributions

Picture: 10

Photo:Unsplash.com

UNIT 1

Binomial Distributions

Introduction

Suppose we consider repeated and independent trials of an experiment with two outcomes; we call one of the outcome SUCCESS and the other outcome FAILURE. Let P be the probability of success. So that q = 1-P is the probability of failure.

Learning Outcomes

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

1. Know different probability distributions that we have, how and when to use each of them.
2. Mention the properties of Binomial distribution
3. Know the mean and variance of Binomial distribution
4. Know how we derive the mean and variance of Binomial distribution

Main Content

8 mins

You should know that if we have a random variable that has only finitely many outcomes, then we can make a table that shows the values *x* in one column and the corresponding probabilities in another column. Such a table is called a probability distribution and is very similar to a relatively frequency distribution. In particular a discrete probability distribution function has two key characteristics:

1. Each probability is between zero and one, inclusive.
2. The sum of the probabilities is one.

The **expected value** or **mean** of a random variable can be thought of as the "long-term" average of the variable. That is, if we performed the probability experiment over and over, we would expect the average of the numerical values to be this amount.

One of the most important types of random variables is the binomial, which arises as follows. Suppose that n independent sub experiments (or trials) are performed, each of which results in either a “success” with probability p or a “failure” with probability 1 − p. If X is the total number of successes that occur in n trials, then X is said to be a binomial random variable with parameters n and p. Before presenting the general formula for the probability that a binomial random variable X takes on each of its possible values 0,1, ..., n, we consider a special case. Suppose that n = 3 and that we are interested in the probability that X is equal to 2. That is, we are interested in the probability that 3 independent trials, each of which is a success with probability p, will result in a total of 2 successes. To determine this probability, consider all the outcomes that give rise to exactly 2 successes:

(s, s, f),(s, f , s),(f , s, s)

The outcome (s, f , s) means, for instance, that the first trial is a success, the second a failure, and the third a success. Now, by the assumed independence of the trials, it follows that each of these outcomes has probability p2(1 − p). For instance, if Si is the event that trial i is a success and Fi is the event that trial i is a failure, then

$$P(s, f, s) = P(S1 \cap F2 \cap S3)$$
$$= P(S1)P(F2)P(S3) \text{ by independence}$$
$$= p(1 - p)p$$

Since each of the 3 outcomes that result in a total of 2 successes consists of 2 successes and 1 failure, it follows in a similar fashion that each occurs with probability).1(2*pp*- Therefore, the probability of a total of 2 successes in the 3 trials is).1(32*pp*- Consider now the general case in which we have n independent trials. Let X denote the number of successes. To determine P{X = i}, consider any outcome that results in a total of i successes. Since this outcome will have a total of i successes and n − i failures, it follows from the independence of the trials that its probability will be *inipp*--)1(. That is, each outcome that results in X = i will have the same probability*inipp*--)1(. Therefore, P{X = i} is equal to this common probability multiplied by the number of different outcomes that result in is successes. Now, it can be shown that there are n!/[i!(n − i)!] different outcomes that result in a total of i successes and n−i failures, where n! (read “n factorial”) is equal to 1 when n = 0 and is equal to the product of the natural numbers from 1 to n otherwise.

That is,

$$0! = 1$$
$$n! = n \cdot (n - 1) \cdots 3 \cdot 2 \cdot 1 \text{ if } n > 0$$

A binomial random variable with parameters n and p represents the number of successes in n independent trials, when each trial is a success with probability p.

If X is such a random variable, then for i = 0,... , n,

$$P(X = i) = \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}$$

The **expected value** or **mean** of a random variable can be thought of as the "long-term" average of the variable. That is, if we performed the probability experiment over and over, we would expect the average of the numerical values to be this amount.

Binomial distribution is a special discrete probability distribution. There are **four conditions** that the experiment has to meet to be considered a binomial experiment:

Conditions for Binomial Experiment

(1) There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.

(2) There are only two possible outcomes, called "success" and "failure," for each trial.

(3) The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial.

(4) The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial, so $p + q = 1$. Since the trials are independent, p remains the same for each trial.

Notation and Properties of the Binomial distribution

(1) $X \sim B(n, p)$ This notation states the random variable X is a binomial distribution within trials and the probability of success , p.

(2) $P(X = k) = \binom{n}{k} P^k q^{n-k}$ Binomial formula

(3) The **mean** of a binomial probability distribution, $\mu = n \cdot p$, and

(4) **Variance** of a binomial probability distribution, $\sigma 2 = n \cdot p \cdot q$.

(5) The **standard deviation**, $\sigma = \sqrt{npq}$

(6) The experiment consists of n repeated and identical trials.

(7) Each trial results in one of two outcomes: one of the outcomes considered "success" and the other "failure".

(8) The repeated and identical trials are independent.

(9) We are interested in x, which represents the total number of successes among the n trials.

Illustration

Example 1

A trainer is teaching a dolphin to do tricks. The probability that the dolphin successfully performs
the trick is 35%, and the probability that the dolphin does not successfully perform the trick is 65%. Out of 10 attempts, you want to find the probability that the dolphin succeeds at most 5 times. State the probability question mathematically.

Solution

Here, if you define X as the number of successful performances, then X takes on the values 0, 1, 2, 3, ..., 10. The probability of success is p = .35. The probability question can be stated mathematically as $P(x \leq 5)$.

$$P(X = 5) = \binom{10}{5}(0.35)^1(0.65)^4 + \binom{10}{5}(0.35)^2(0.65)^3 + \ldots + \binom{10}{5}(0.35)^4(0.65)^1$$

$P(X \leq 5) = 0.905$

Example 2

A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than ten heads?

Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, $p = 0.5$ and $q = 0.5$. The number of trials is $n = 15$. State the probability question mathematically. Find the probability.

Solution

$P(x > 10) = P(x = 11) + P(x = 12) + P(x = 13) + P(x = 14) + P(x = 15)$

$= 1 - P(\text{complement})$

$= 1 - P(x \leq 10)$

$= 1 - b(15, .5, 10)$

$$= 1 - \binom{15}{11}(0.5)^{11}(0.5)^{15-11} + \binom{15}{12}(0.5)^{12}(0.5)^3 + \ldots + \binom{15}{15}(0.5)^{15}(0.5)^0$$

$= 1 - 0.941$

$= 0.059$

Example 3

Three fair coins are flipped. If the outcomes are independent, determine the
probability that there are a total of i heads, for i = 0, 1, 2, 3.

Solution

If we let X denote the number of heads ("successes"), then X is a binomial
random variable with parameters n = 3, p = 0.5. By the preceding we have

$$P(X=0)=\binom{3}{0}(0.5)^0(0.5)^3=\frac{1}{8}$$

$$P(X=1)=\binom{3}{1}(0.5)^1(0.5)^2=\frac{3}{8}$$

$$P(X=2)=\binom{3}{0}(0.5)^2(0.5)^1=\frac{3}{8}$$

$$P(X=3)=\binom{3}{3}(0.5)^3(0.5)^0=\frac{1}{8}$$

Example 4

(a) Determine P{X ≤ 12} when X is a binomial random variable with parameters 20 and 0.4.

(b) Determine P{Y ≤ 10} when Y is a binomial random variable with parameters 16 and 0.5.

Solution

(a) P{X ≤ 12} = 0.9790

(b) P{Y ≤ 10} = 1 − P{Y < 10} = 1 − P{Y ≤ 9} = 1 − 0.7728 = 0.2272

Example 5

Suppose that a particular trait (such as eye color or handedness) is determined by a single pair of genes, and suppose that d represents a dominant gene and r a recessive gene. A person with the pair of genes (d, d) is said to be pure dominant, one with the pair (r, r) is said to be pure recessive, and one with the pair (d, r) is said to be hybrid. The pure dominant and the hybrid are alike in appearance. When two individuals mate, the resulting offspring receives one gene from each parent, and this gene is equally likely to be either of the parent's two genes. What is the probability that the offspring of two hybrid parents has the opposite (recessive) appearance? Suppose two hybrid parents have 4 offsprings. What is the probability 1 of the 4 offspring has the recessive appearance?

Solution

The offspring will have the recessive appearance if it receives a recessive gene from each parent. By independence, the probability of this is (1/2)(1/2) = 1/4. Assuming the genes obtained by the different offspring are independent (which is the common assumption in genetics), it follows from part (a) that the number of offspring having the recessive appearance is a binomial random variable with parameters n = 4 and p = 1/4. Therefore, if X is the number of offspring that have the recessive appearance, then

$$P(X=1)=\binom{4}{3}\binom{1}{4}^1\binom{1}{4}^3$$

$$=4\binom{1}{4}^1\binom{1}{4}^3$$

$$=\frac{27}{64}$$

Example 6

If a ludo die is thrown five times, calculate the probability of getting each of the following outcomes;

(a) Exactly one six

(b) Exactly two sixes

(c) Three or more sixes

Solution

In this situation, the binomial parameters are defined as follows;

n = 5

p = P(success) = P(six) = 61

q = (1 – p) = (1 - 61) = 65

Since there are six possible outcomes per trial and only one of the can result to a “six”. Success in this case is the occurrence of a “six”. Hence, x is the appearance of a “six”.

(a) $P(X = x)\binom{n}{x}p^x q^{n-x}$

$$P(\text{exactly one six}) = P(X = 1) = \binom{5}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4 = 0.4019$$

(b) $P(X = x)\binom{n}{x}p^x q^{n-x}$

$$P(\text{exactly two sixes}) = P(X = 2) = \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = 0.1608$$

(c) $P(\text{three or more sixes}) = P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5)$

$$P(X = 3) = \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = 0.0322$$

$$P(X = 4) = \binom{5}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^1 = 0.0032$$

$$P(X = 5) = \binom{5}{5}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^0 = 0.0001$$

$$P(\text{three or more sixes}) = 0.0322 + 0.0032 + 0.0001 = 0.0355$$

Summary

In this unit you have learnt:

- How to obtain values for binomial distribution and
- That the value obtained must be between 0 and 1.

Self Assessment Questions

(1) Know different probability distributions that we have, how and when to use each of them.
(2) Mention the properties of Binomial distribution
(3) Know the mean and variance of Binomial distribution
(4) Know how we derive the mean and variance of Binomial distribution

Tutor Marked Assessment

- If a fair coin is tossed six times, find the probability of obtaining;
- Exactly six heads
- Exactly four heads
- At least two heads
- The probability that a patient recovers from a myocardial infarction (heart attack) is 0.4. If 15 people are known to have contacted the disease, find the probability that;
- At least ten will survive
- Between 3 to 10 will survive
- Exactly 5 will survive
- A pharmaceutical company manufactures syringes which are sent out to customers in lots of 10,000. The company operates an acceptance sampling scheme; whereby a random sample of 10 syringes is taken from each lot ready for dispatch. The lot is released only if the number of defective syringes in the sample is less than 3, otherwise the whole lot of 10,000 is rejected and reprocessed.
- If 10% of all syringes produced are known to be defective, find the proportion of lot that will be rejected.
- If the company replaces all its syringes producing machines causing the proportion of defective syringes in the sample of 10 is less than 2 ; will be proportion of lots rejected decrease? If so, find the number of syringes per lot the company expect to save by this replacement.
- The probability that Bill hits a target is P = 1/5. He fires 100 times. Find the expected number μ of times he will hit the target and the standard deviation (σ).

Reference

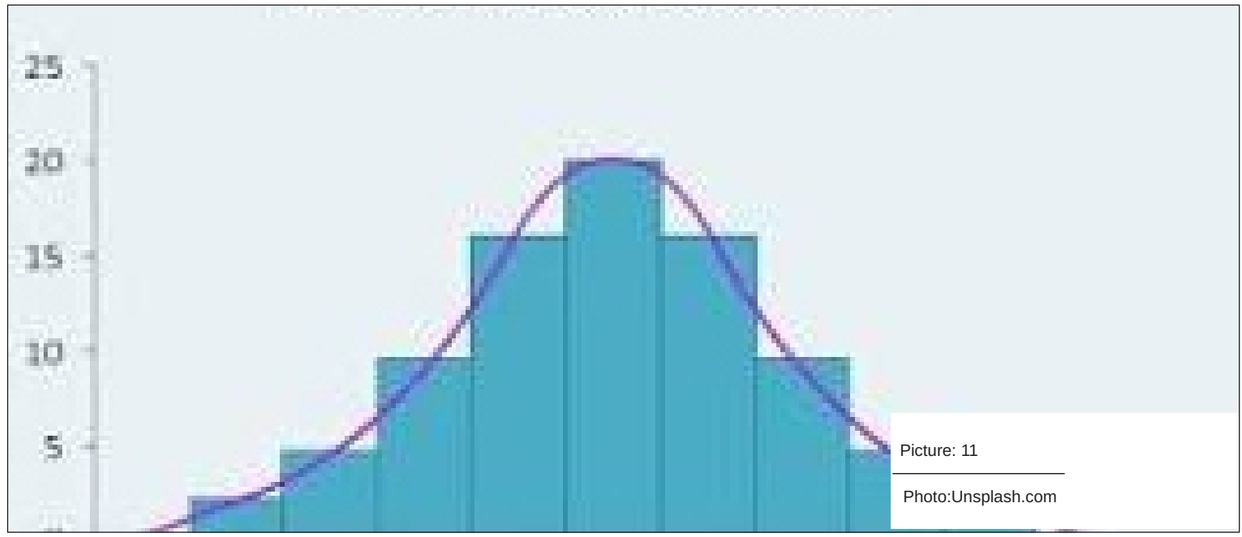
Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.

Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.

Gupta (2014). Introduction to Fundamental Statistics.

Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.

Introductory Statistics 3rd Edition by Sheldom M. Ross.



Picture: 11

Photo:Unsplash.com

UNIT 2

Continues Binomial Distributions

Introduction

Any experiment which consists of repeated trials, each with two possible outcomes, which may be labeled “success” or “failure” and if the repeated trials are independent and the probability of a success remains constant from trial to trial, then such an experiment is known as binomial experiments. Binomial distribution can also be defined as the sum of independent Bernoulli random variables. A random X with probability mass function P(X = 0) = 1- p and P(X = 1) = p for (0)1ff p is said to have a Bernoulli distribution. If X_1, X_2 ...Xn are n independent Bernoulli random variables, then the sum of these n independent Bernoulli random variables is a binomial distribution.

Learning Outcomes

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

1. Know different probability distributions that we have.
2. Express the Binomial distribution.
3. Know the importance of Binomial distribution.

Main Content

5 mins

It will interest you to know that the Binomial Distribution was so named by the Scottish statistician George Undy Yule in 1911. The binomial formula was first described in 1676 by Sir, Isaac Newton (1642 – 1727), the great English mathematician and physicist. It is proof, for positive integer exponents, was given by the Swiss mathematician, Jacob Bernoulli (1654 – 1705) in a 1713 publication. Each observed event from binomial distribution is sometimes called a "Bernoulli trial".

In general, suppose we have a random sequence in which the outcomes of each individual trial is one of two types A or B those outcomes occurring with probabilities P and (1 – P), respectively. Consider a group of n observations from this random sequence. It will be convenient to refer to each such group as a sample of n observations. We can comfortably find the probability distribution of the number of A's in the sample. This number we shall call x, and clearly x must be one of one numbers 0, 1, 2, …,n. We shall also define ;

$p = \frac{x}{n}$, the proportion of as in the sample

$q = \frac{n-x}{n} = 1 - p$, the proportion of Bs in the sample.

We argue that the probability of x A's and (n – x) Bs is $P^x (1 - P)^{n-x}$ multiplied by the number of ways in which one can choose x out of n sample members to receive a label 'A'. This multiplying factor is called a binomial coefficient. In simple cases, we work out by simple enumeration, but clearly this could be tedious with large values of n and x. As discussed earlier, the binomial coefficient is usually denoted by the following:

$\binom{n}{x}$ Called n combination x

Where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

Hence, the probability that the sample of n individuals contains x A's and (n – x) B's is given by the binomial probability mass function shown below:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

n is the number of trials

x is the number of success

x = 0, 1, 2, …, n

P is the probability of success

q = 1 – p is the probability of failure

$\binom{n}{x} = \frac{n!}{(n-x)!x!}$

Where $n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$

Which is called n factorial and by definition 0! = 1

Note that it is important that 'Success' in this case refers to the outcomes in favour of the experiment under study; it does not take the literal meaning.

The binomial distribution can also be defined as the sum of independent Bernoulli random variables. A random X with probability mass function P(X = 0) = 1 – P and P(X = 1) = P for $(0 \le p \le 1)$ is said to have a Bernoulli distribution. If $X_1, X_2, \ldots, X_n$ are n independent Bernoulli random variables, then the sum of these n independent Bernoulli random variables is a binomial distribution. That is:

If $x_1, x_2, \ldots, x_n$ be a random sample of size n from Ber(P), which has mean μ = P and variance $\sigma^2 = P(1 - p)$; the $\sum_{i=1}^{n} X_i$ is B(n, p).

Illustration

Example 1

If a die is tossed five times, find the mean and variance of obtaining a 'six' in this experiment.

Solution

n = 5 (sample size)

P = 1/6 (probability of success)

q = 5/6 (probability of failure)

mean = μ= np = 5x1/6 = 0.8333

variance = σ^2 = npq = 5 x 1/6 x 5/6 = 0.6944

Example 2

A box contains two black balls out of a total of seven. A ball is selected at random, its colour noted, the ball is then replaced and the box shaken. If 50 balls are selected in this manner; find the mean of the number of black balls and the variance.

Solution

n = 50

P = 2/7 (probability of success)

q = 5/7 (probability of failure)

Mean = μ= np = 50 x 2/7 = 14.290

Variance = σ^2 = npq = 50 x 2/7 x 5/7 = 10.2041

Example 3

Suppose you play a game that you can only either win or lose. The probability that you win any game is 55%, and the probability that you lose is 45%. Each game you play is independent.

(a) If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times.

(b) Find the mean number of wins.

(c) Find the standard deviation of wins

Solution

(a) Let X be the number of wins (0, 1, 2, 3, ..., 20). The probability of a success is p =0.55.

The number of trials is n = 20. The probability question can be stated mathematically as

$$P(X=15)=\binom{20}{15}(0.55)^{15}(0.45)^{20-15}$$

$$= 0.0365$$

(b) μ = $n \cdot p$ = 20·.55 = 11 wins is the mean number of wins for 20 trials.

(c)σ = $\sqrt{npq} = 2.22$

Example 4

A fair coin is tossed 6 times call heads a success. This is a binomial experiment with n = 6, and p = q = 0.5

(a)The probability that exactly two heads occurs.

(b)The probability of getting at least four heads

(c) The probability of getting no heads

(a) $P(2)=\binom{6}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^4=\frac{15}{64}=0.230$

(b) P(4) + P(5) + P(6) = -

$$=\frac{15}{64}+\frac{6}{64}+\frac{1}{64}=\frac{11}{32}=0.260$$

(c) $1-q^n=1-64=1-\frac{1}{64}=\frac{63}{64}=0.980$

Example 5

Suppose 20% of the items produced by a factory are defective, Suppose 4 items are chosen at random. Find the probability that:

(a)2 are defective

(b)3 are defective

None are defective

Solution

This is a binomial experiment with n =4, p = 0.2 and q = 1 – p = 0.8; that is B(4,0.2).

Hence:

(a) K =2 and $P(2)=\binom{4}{2}(0.2)^2(0.8)^2=0.1536$

(b) K = 3 and $P(3)=\binom{4}{3}(0.2)^3(0.8)=0.0256$

(c) $P(0)=q^4=(0.8)^4=0.4095$

Hence P(X>0) = 1 – P (0) = 1 - 0.4095 = 0.5904

Example 7

A fair coin tossed 6 times call head a success. This is a binomial experiment with n = 6 and p =q = 0.5

Solution

The probability that exactly two heads occurs

$$P(2) = \binom{6}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^4 = \frac{15}{64} = 0.230$$

$$P(4) + P(5) + P(6) = \binom{6}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^2 + \binom{6}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right) + \binom{6}{6}\left(\frac{1}{2}\right)^6$$

$$= \frac{15}{64} + \frac{6}{64} + \frac{1}{64} = \frac{11}{32} = 0.340$$

Example 8

The probability that Ann hits a target is p = 1 /3 hence she misses with probability q = 1 – p = 2/3. She fire seven times. Find the probability that she hits the target:

(a) Exactly 3 times

(b) At least one time

Solution

(a) K = 3 nhence the probability that she hits the target three times is

$$P(3) = \binom{7}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^4 = \frac{560}{2187} = 0.260$$

(b) The probability that she never hits the target that is, all failures is $q^7 = \left(\frac{2}{3}\right)^7 = \frac{128}{2187} = 0.06$.

Thus the probability that she hits the target at least once is $1 - q^7 =$

$$\frac{2059}{2187} = 0.940 = 94 \text{ percent.}$$

Summary

In this unit you have learnt:

- How to obtain values for binomial distribution and
- That the value obtained must be between 0 and 1.

Self Assessment Questions

(1) Know different probability distributions that we have.

(2) Express the Binomial distribution.

(3) Know the importance of Binomial distribution.

Tutor Marked Assessment

- In a large town, one person in ten is red haired, you are required to find:
- The probability that a random sample of 20 persons will contain at least 4 red haired persons.
- Compute P(k) for the binomial distribution B(n,p) where;

$n = 5, P = \frac{1}{4}, k = 2$

$n = 10, P = \frac{1}{2}, k = 7$

$n = 8, p = \frac{2}{3}, k = 5$

- The mathematics department has eight graduate assistants who are assigned the same office. Each assistant is just as likely to study at home as in the office. Find the minimum number m of desks that should be put in the office so that each assistant has a desk at least 90% of the time.

  Determine the expected number of girls if male and female children are equally probable.
- Find the probability p that the expected number of girls does occur.
- The probability that a man hits a target is p =0.1. He fires n = 100 times. Find the expected number E of times he will hit the target and the standard deviation.
- A student takes an 18 questions multiple – choice exam, with four choices per question. Suppose one of the choices is obviously incorrect, and the student make an educated guess of the remaining
- choices. Find the expected number E of correct answer and the standard deviation.
- The probability that John hits a target is p = 1 /4. He fires n = 6 times.
- Find the probability that he hits the target:

  Exactly 2 times (b) More than four times (c) At least once.

References

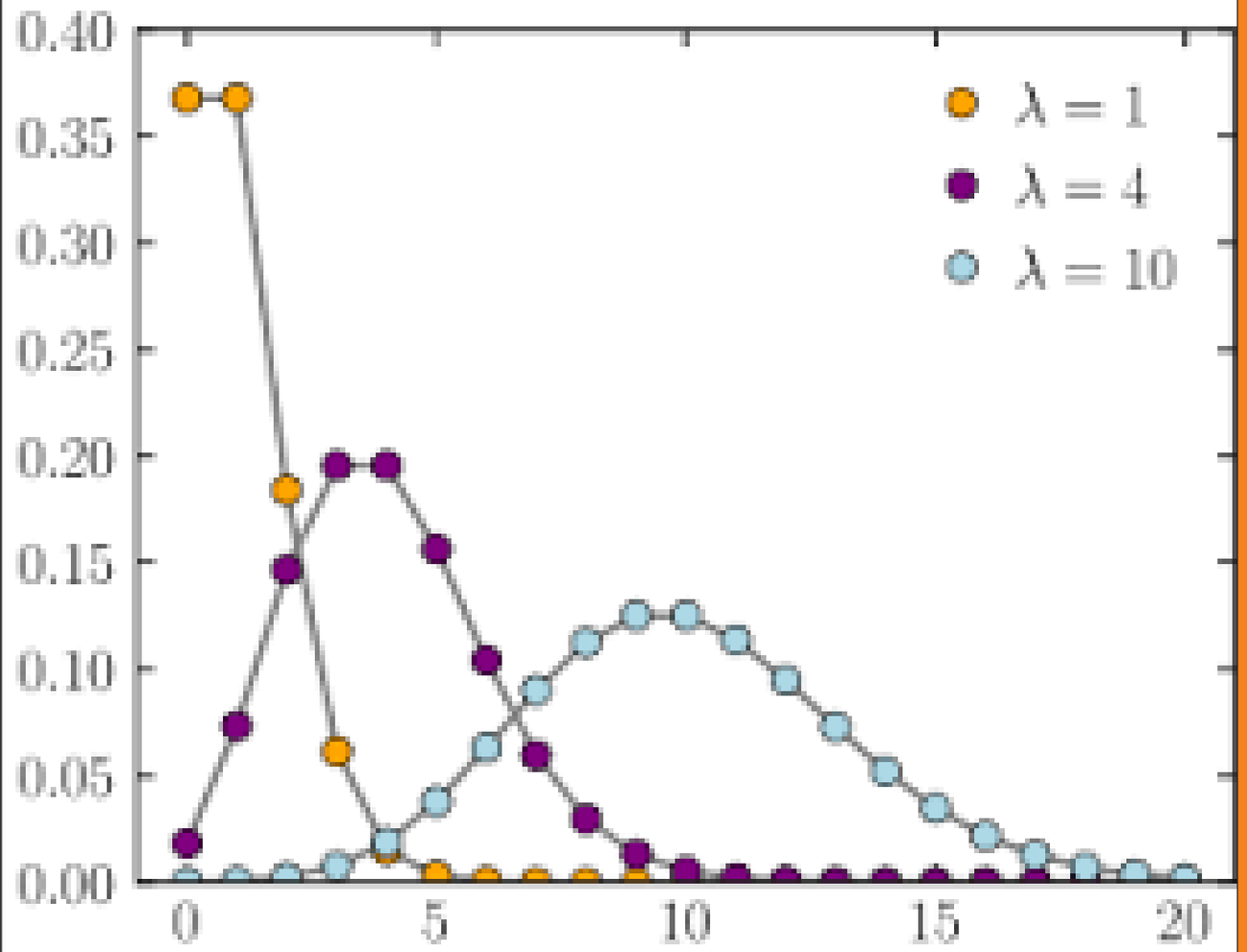
Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.

Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.

Gupta (2014). Introduction to Fundamental Statistics.

Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.

Introductory Statistics 3rd Edition by Sheldom M. Ross.

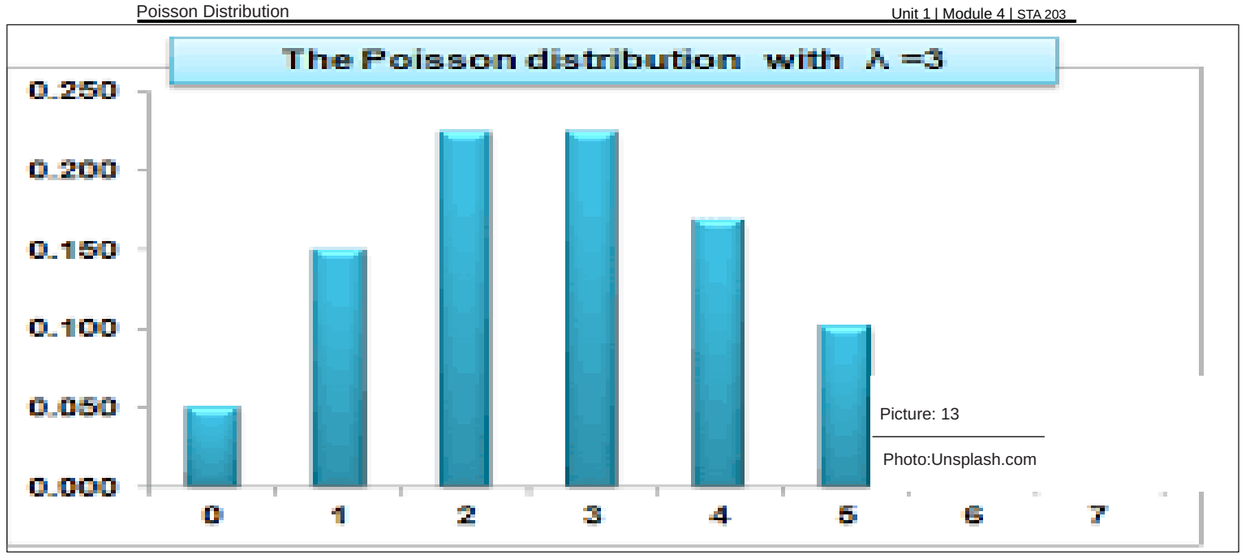


Picture: 12

Photo:Unsplash.com

Module 4

Poisson Distributions



Picture: 13

Photo:Unsplash.com

UNIT 1

Poisson Distributions

Introduction

The Poisson probability distribution is sometimes useful as a limiting form of the binomial, but it is important also in its own right as a distribution arising when events of some sort occur randomly in time or when small particles are distributed randomly in space.

Learning Outcomes

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

1. Know the Poisson distributions that are useful in their field of study.
2. Compute the Poisson distribution
3. Relate the importance of Poisson the discrete random variable.

Main Content

6 mins

Here, we will use upper case letters such as *X* or *Y* to denote a random variable. Lower case letters like *x* or *y* denote the value of a random variable. If *X* is a random variable, then *X* is described in words, and the value *x* is given as a number. If we have a random variable that has only finitely many outcomes, then we can make a table that shows the values *x* in one column and the corresponding probabilities in another column. Such a table is called a *probability distribution* and is very similar to a relatively frequency distribution. In particular, a discrete probability distribution function has two key characteristics which are:

1. Each probability is between zero and one, inclusive.
2. The sum of the probabilities is one.

Both Binomial and Poisson are discrete probability distributions.

In Binomial the goal was to look for the probability of a specific value of success In *n* trials. Now we want to look for the specific number of occurrences in a specific amount of time or space. A random variable X that takes on one of the values 0, 1, 2,… is said to be a Poisson random variable with parameter λ if for some positive value λ its probabilities are given by

$$P(X=i)=\frac{C\lambda^{i}}{i!} \quad i=0,1,\ldots\ldots$$

In the preceding c is a constant that depends on l. Its explicit value is given by c = e−λ, where e is a famous mathematical constant that is approximately equal to 2.718. A random variable X is called a Poisson random variable with parameter λ if

$$P(X=i)=\frac{e^{-\lambda}\lambda^{i}}{i!} \quad i=0,1,\ldots\ldots$$

A discrete random variable X is said to have the Poisson distribution with parameter $\lambda>0$ if X takes on non negative integer values k = 0, 1, 2,…… with respective probabilities

$$P(k)=f(k;\lambda)=\frac{e^{-\lambda}\lambda^{k}}{k!}$$

Such a distribution will be denoted by POI(λ). This distribution named after Simeon Poisson (1781 – 1840) who discovered it in early part of the 19th century.

The Poisson distribution appears in many natural phenomena, such as the number of telephone calls per minute at some switchboard, the number of misprints per page in a large text, and the number of α particles emitted by a radioactive substance. Although the Poisson distribution is of independent interest, it also provides us with a close approximation of the binomial for small k provided p is small and λ = np more specifically, if n ³ 50 and np < 5. This property is indicated in the above which compares the binomial and Poisson distributions for small values of k with n = 100, p = 1/100, and λ = np = 1.

Conditions of Poisson Experiment

Note that the following are conditions of Poisson experiment

(1) The experiment consists of counting the number of events occurring in a fixed interval of time or space if these events happen with a known average rate and independently of the time since the last event.

(2) The probability of the event remains constant for each interval of equal length.

(3) The number of occurrences in one fixed interval is independent of the number of occurrences in of interest

Now, let us consider other fixed intervals.

The random variable X = the number of occurrences in the interval of interest. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages.

Properties for the Poisson Distribution

(1) X ~ P() where is the mean number of occurrences per fixed interval.

(2) μ = np

(3) σ = m

(4) The probability of exactly x occurrences in an interval is

$$P(x)=\frac{\mu^{x}e^{-\mu}}{x!}$$

Illustration

Some examples are given below:

Example 1

If X is a Poisson random variable with parameter λ = 2, find P{X = 0}.

Solution

$$P(X=0)=\frac{2^0 e^{-2}}{0!}=0.1353$$

Example 2

Suppose that items produced by a certain machine are independently defective w ith probability 0.1. What is the probability that a sample of 10 items will contain at most 1 defective item? What is the Poisson approximation for this probability?

Solution

If we let X denote the number of defective items, then X is a binomial random variable with parameters n = 10, p = 0.1. Thus the desired probability is

$$P(X=0)+P(X=1)=\binom{10}{0}(0.1)^0(0.9)^{10}+\binom{10}{1}(0.1)^1(0.9)^9=0.7361$$

Since np = 10(0.1) = 1, the Poisson approximation yields the value

$$P(X=0)+P(X=1)=e^{-1}+e^{-1}=0.7358$$

Thus, even in this case, where n is equal to 10 (which is not that large) and p is equal to 0.1 (which is not that small), the Poisson approximation to the binomial probability is quite accurate.

Both the expected value and the variance of a Poisson random variable are equal to λ. That is, we have the following.

If X is a Poisson random variable with parameter λ,λ > 0, then

E[X] = λ

Var(X) = λ

Example 3

Suppose the average number of accidents occurring weekly on a particular highway is equal to 1.2. Approximate the probability that there is at least one accident this week.

Solution

Let X denotes the number of accidents. Because it is reasonable to suppose that there are a large number of cars passing along the highway, each having a small probability of being involved in an accident, the number of such accidents should be approximately a Poisson random variable. That is, if X denotes the number of accidents that will occur this week, then X is approximately a Poisson random variable with mean value λ = 1.2. The desired probability is now obtained as follows:

$$P(X>0)=1-P(X=0)$$

$$=\frac{1-e^{-1.2}(1.2)^0}{0!}$$

$$=1-e^{-1..2}$$

$$=1-0.3012$$

$$=0.6988$$

Therefore, there is approximately a 70 percent chance that there will be at least one accident this week.

Example 4

For the Poisson distribution $f(k;\lambda)=\frac{\lambda^k e^{-\lambda}}{k!}$ Find $f(2,1)$

Solution

$$f(2,1)=\frac{1^2 e^{-1}}{2!}=\frac{0.368}{2}=0.273$$

Example 5

For the Poisson distribution $f(k;\lambda)=\frac{\lambda^k e^{-\lambda}}{k!}$

Solution

$$f(3,\frac{1}{2})=\frac{\left(\frac{1}{2}\right)^3 e^{-0.5}}{3!}=\frac{e^{-0.5}}{48}=\frac{0.607}{48}=0.013$$

Example 6

For the Poisson distribution $f(k;\lambda)=\frac{\lambda^k e^{-\lambda}}{k!}$ Find $f(2,0.7)$

Solution

$$f(2,0.7)=\frac{(0.7)^2 e^{-0.7}}{2!}=\frac{(0.49)(0.497)}{2}=0.12$$

Example 7

If X is a Poisson random variable with parameter λ = 2, find P{X = 0}.

Solution

$$P(X=0)=\frac{e^{-2}2^0}{0!}$$

Using the facts that 2^0 = 1 and 0!= 1, we obtain

P{X = 0} = 2-*e* = 0.1353

In the preceding, the value of 2-*e* was obtained from a table of exponentials. Alternatively, it could have been obtained from a scientific hand calculator or a personal computer.

Poisson random variables arise as approximations to binomial random variables. Consider n independent trials, each of which results in either a success with probability p or a failure with probability 1 − p. If the number of trials is large and the probability of a success on a trial is small, then the total number of successes will be approximately a Poisson random variable with parameter λ = np.

Some examples of random variables whose probabilities are approximately given, for some λ, by Poisson probabilities are the following:

(1) The number of misprints on a page of a book

(2) The number of people in a community who are at least 100 years old

(3) The number of people entering a post office on a given day

Each of these is approximately Poisson because of the Poisson approximation to the binomial. For instance, we can suppose that each letter typed on a page has a small probability of being a misprint, and so the number of misprints on a page will be approximately a Poisson random variable with parameter λ = np, where n is the large number of letters on a page and p is the small probability that any given letter is a misprint.

Summary

In this unit you have learnt:

- How to obtain values for binomial distribution and
- That the value obtained must be between 0 and 1.

Self Assessment Questions

(1) Know the Poisson distributions that are useful in their field of study.

(2) Compute the Poisson distribution

(3) Relate the importance of Poisson the discrete random variable.

Tutor Marked Assessment

- An electronic company manufactures a specific components for an ultrasound machine. It is founds that out of every 1000 components produced, 8 are defective. If the components are packed in batches of 250 find the probability of obtaining no defective in a batch.
- An electronic company manufactures a specific components for an ultrasound machine. It is founds that out of every 1000 components produced, 8 are defective. If the components are packed in batches of 250 find the probability of obtaining exactly defective in a batch.
- An electronic company manufactures a specific components for an ultrasound machine. It is founds that out of every 1000 components produced, 8 are defective. If the components are packed in batches of 250 find the probability that the batch contains at least 3 defectives.
- Define a random variable that follows a Poisson distribution with parameter(Θ).
- What are the properties of a Poisson distribution
- Define a random variable that follows a Geometric distribution.

Reference

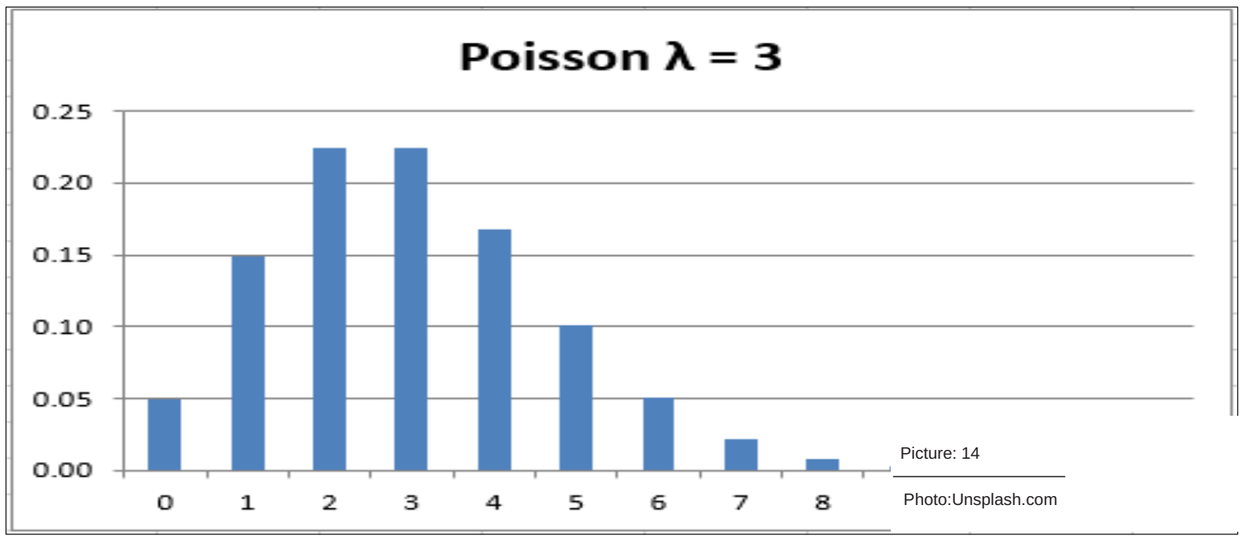
Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and

Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.

Gupta (2014). Introduction to Fundamental Statistics.

Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.

Introductory Statistics 3rd Edition by Sheldom M. Ross



Picture: 14

Photo:Unsplash.com

UNIT 2

Continuation Of Poisson Distributions

Introduction

A random variable X that takes on one of the values 0, 1, 2, . . . is said to be a Poisson random variable with parameter λ and the random variable X is said to follow a Poisson distribution if the probability mass function is given as

$$P(k) = f(k;\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!} \quad , k = 0, 1, 2, \ldots\ldots\ldots\ldots$$

Learning Outcomes

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

1. Know the properties of the Poisson distributions.
2. Compute the values for Poisson distribution
3. Relate important of Poisson distribution
4. Prepare the mean and variance of Poisson distribution

Main Content

5 mins

It will interest you to know that, Abraham de Moivre (1667 – 1754) has apparently described in previously in 1718. It was also described independently by others, including William Sealy Gosset in 1909.

The Poisson probability distribution is sometimes useful as a limiting form of the binomial, butn it is important also in its own right as a distribution arising when events of some sort occur randomly in time, or when small particles are distributed randomly in space. A very good instance of this probability model is that of the emission of radioactive particles from some radioactive materials. The rate of emission l will be constant but the particles will be emitted in a purely random way. The situation is therefore almost exactly the same as a sequence of n binomial trials. Where in each of the trials there is a probability *n*l of there being an event and (1 -) of there being no event. The probability that the whole series of n trials provides exactly x events is, in this approximation, given by the binomial distribution:

$$\frac{n(n-1)...(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^{x}\left(1-\frac{\lambda}{n}\right)^{n-x}$$

Now, this binomial approximation will get better and better as n increases and we can replace n(n – 1)...(n – x + 1) by n^x since x will be negligible in comparison with n.

Similarly, we can replace $(1 - \lambda/n)^{n-x}$ by $(1 - \lambda/n)^{n}$ since $(1 - \lambda/n)^{x}$ will approach 1 as n increases. It is a standard mathematical result that, as n increases indefinitely, $(1 - \lambda/n)^{n}$ approaches $e^{-\lambda}$, where e is the base of natural (or Napierian)logarithms (e = 2.718….). Finally, then, in the limit as n increases indefinitely, the probability of x events approaches:

$$P(X)=\frac{n^x}{x!}\left(\frac{\lambda}{n}\right)^{x}e^{-\lambda}=\frac{e^{-\lambda}\lambda^{x}}{x!}$$

Hence, Poisson distribution is used as a limiting distribution to the binomial distribution when n becomes large and p is fairly small. In most cases, when n^3 30 and p < 0.01, the calculation of various probabilities using the binomial formula becomes tedious; hence the Poisson distribution could serve as a good approximation to the binomial distribution. When a random variable X follows the Poisson distribution with parameter l; it can easily be written as)(~lPX.

Properties of the Poisson Distribution

You should note that every Poisson distribution should possess the following properties:

(1) The average number of success l occurring in the given time interval or specified region is known.

(2) The probability that a single success will occur during a very short time interval or in a small region is proportional to the length of the time interval or the size of the region and does not depend on the number of successes occurring outside this time interval or region.

(3) The probability that more than one success will occur in such a short time interval or falling in such a small region is negligible.

The Mean and Variance of the Poisson Distribution

The Poisson distribution is determined entirely by the one parameter It follows that all the features of the distribution in which one might be interested are functions of only . The expectation and variance of a Poisson random variable X can simply be obtained by applying the general formulae of expectation and variance to the probability distribution and using standard algebraic results on summation of series:

$$E(X)=\sum_{x=0}^{\infty}x.p(X=x)=\sum_{x=0}^{\infty}x^{2}.\frac{e^{-\lambda}\lambda^{x}}{x!}=\lambda$$

This result following after a little algebraic manipulation; by similar manipulations we find the following results.

Recursive Formula Of The Poisson Distribution

The Poisson recursive formula enables us to compare, more easily, the successive probabilities of a Poisson distribution from earlier computed ones or existing probabilities. The recursive formula is also known as the recurrence relations. For example, with the recursive formula, one can easily compute P(X = 2) from P(X = 1) and so forth. The Poisson recursive formula is given by:

$$P(X=x+1)=\frac{\lambda}{x+1}\times P(X=x)\quad \text{, for x = 1, 2, ...}$$

$$V(X)=E(X^{2})-\left[E(X)\right]^{2}=\lambda$$

Thus, the variance of x, like the mean, is equal to l and the standard deviation is therefore $\sqrt{\lambda}$

If a random variable X follows the Poisson distribution with parameter l its mean and variance are respectively given as follows:

$\mu = E(X) = \lambda$

$\sigma^2 = V(X) = \lambda$

Illustration

Some examples are given below:

Example 1

If X ~P(1.6)

Find P(X=0)

Solution

$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for x = 0,1,2,……

Since $X \sim P(1.6)$ then; $P(X = x) = \frac{e^{-1.6}(1.6)^x}{x!}$

(1) $P(X = 0) = \frac{e^{-1.6}(1.6)^0}{0!} = e^{-1.6} = 0.2019$

Example 2

If X ~P(1.6)

Find P(X=1)

Solution

$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for x = 0,1,2,……

Since $X \sim P(1.6)$ then; $P(X = x) = \frac{e^{-1.6}(1.6)^x}{x!}$

$P(X = 1) = \frac{e^{-1.6}(1.6)^1}{1!} = 1.6 \times 0.2019 = 0.3230$

Example 3

If X ~P(1.6)

Find P(X=2)

Solution

$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for x = 0,1,2,……

Since $X \sim P(1.6)$ then; $P(X = x) = \frac{e^{-1.6}(1.6)^x}{x!}$

$P(X = 2) = \frac{e^{-1.6}(1.6)^2}{2!} = \frac{0.2019 \times (1.6)^2}{2} = 0.2584$

Example 4

An electronic company manufactures a specific component for an ultrasound machine. It finds that out of every 1000 components produced, 8 are defective. If the components are packed in batches of 250 find

(a) The probability of obtaining no defective in a batch

(b) The probability of obtaining exactly two defectives in a batch

(c) The probability that the batch contains at least 3 defective.

Solution

$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for x = 0, 1, 2, …..

In this case probability of a defective is $p = \frac{8}{1000} = 0.008$

n = 250

$\lambda = np = 250 \times 0.008 = 2$

X ~ P(2)

$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$

(a) P(no defective) = P(X = 0) = $P(X = 0) = \frac{e^{-2}2^0}{0!} = 0.1353$

(b) P(two defective) = P(X = 2) = $P(X = 2) = \frac{e^{-2}2^2}{2!} = 0.2706$

(c) P(at least 3 defectives) = P(X ≥ 3)

= P(X = 3) + P(X = 4) +……= 1 – P(X< 3)

$$P(X < 3) = P(X= 0) + P(X = 1) + P(X =2)$$

$$P(X = 0) = P(X = 0) = \frac{e^{-2}2^{0}}{0!} = 0.1353$$

$$P(X = 1) = P(X = 1) = \frac{e^{-2}2^{1}}{1!} = 0.2706$$

$$P(X = 2) = P(X = 2) = \frac{e^{-2}2^{2}}{2!} = 0.2706$$

$$P(X < 3) = 0.1353 + 0.2706 + 0.2706 = 0.6765$$

P(at least 3 defectives) = 1 – P(X<3) = 1 – 0.6765 = 0.3235

Example 5

The number of deaths by road, as recorded in a country, is 2 per 50,000 of the population. Find the probability in a town of 100,000 inhabitants within the country.

(a) Exactly five deaths by road, will be recorded.

Solution

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad \text{for x = 0, 1,2, ……}$$

In this case probability of death by road is given by

$$p = \frac{2}{50000} = 0.00004$$

n = 100,000

λ = np = 100000 x 0.00004 = 4

X ~ P(4)

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

(a) P(exactly five deaths) = $P(X = 5) = \frac{e^{-4}4^{5}}{5!} = 0.1563$

Summary

In this unit you have learnt:

- How to obtain values for Poisson distribution
- That the value obtained must be between 0 and 1
- How to obtain the expected value and variance.

Self Assessment Questions

(1)Know the properties of the Poisson distributions.

(2)Compute the values for Poisson distribution

(3)Relate important of Poisson distribution

(4)Prepare the mean and variance of Poisson distribution

Tutor Marked Assessment

- If X ~P(1.6)
  Find P(X£2)
- If X ~P(1.6)
  Find P(X< 2)
- Suppose that on the average one gynecologist in 1,000 makes na technical error in conducting the caesarean section operation. If 10,000 cases of caesarean section operations are selected at random and examined; find the probability that between six and eight of the operations would be in error.
- The number of deaths by road, as recorded in a country, is 2 per 50,000 of the population. Find the probability in a town of 100,000 inhabitants within the country:
  Between one and three deaths by road, inclusive, will be recorded
- The mean number of the death by road.
- The variance
- The standard deviation
- If X ~ P(1.8) use the Poisson recursive formula to calculate the probability of 5 or more.
- The number of deaths by road as recorded in a country, is 2 per 50,000 of the population. Find the probability in a town of 100,000
- inhabitants within the country between two to eight deaths by road will be recorded using the Poisson recursive formula.

References

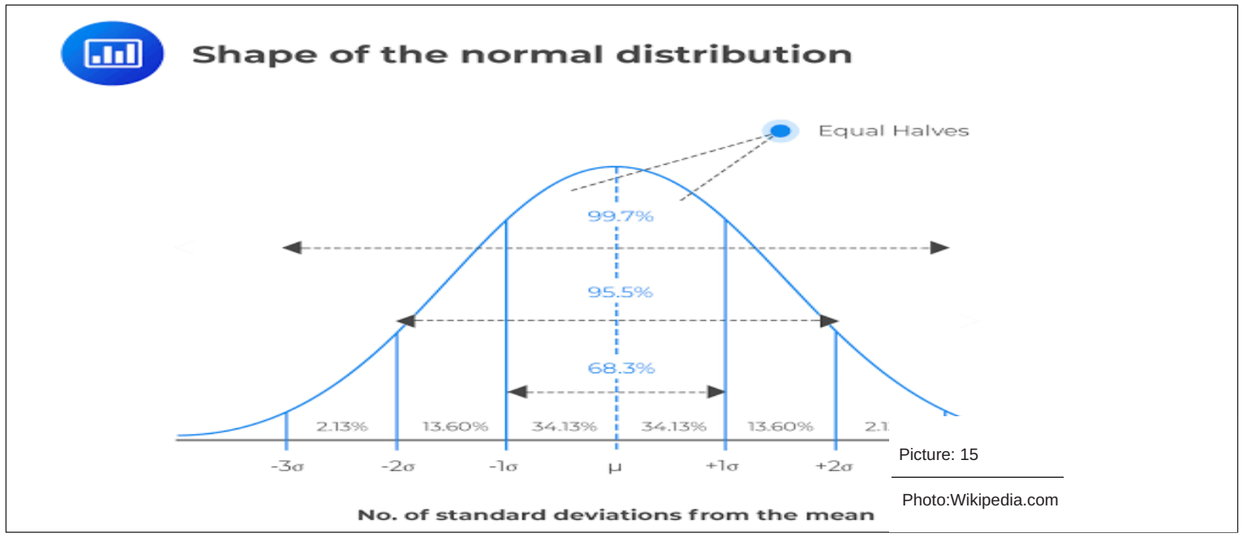
- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and
- Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.
- Gupta (2014). Introduction to Fundamental Statistics.
- Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.
- Introductory Statistics 3rd Edition by Sheldom M. Ross

Picture: 15

Photo:Unsplash.com

Module 5

Normal Distributions



Picture: 15

Photo:Wikipedia.com

UNIT 1

Normal Distributions

Introduction

The most important example of a continuous random variable X is the normal random variable whose density function has a bell-shaped graph. More precisely, there is a normal random variable X for each pair of parameters σ > 0 and μ, where the corresponding densitv function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[\frac{-1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right]$$

Such a normal distribution with parameters m and s will be denoted by),(2smN. If Xis such a continuous random variable, then we say X is normally distributed or that X is .

Learning Outcomes

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

1. Describe the Normal distribution
2. The uses of normal distribution
3. Describe the properties of normal distribution
4. Compute normal distribution

Main Content

6 mins

It may interest you to note that the normal distribution is the most important continuous probability distribution in the field of statistics. The graph of the normal distribution is bell – shaped, called the normal curve. The normal distribution is often attributed to P.S Laplace and C. F Gauss whose name it bears. However, its origin dates back to the works of Jacob Bernoulli in 1713; who provided the first basic elements of the law of large numbers. In 1733 Abraham de Moivre was the first to obtain the normal distribution as an approximation of the binomial distribution. The normal distribution is often referred to as the Gaussian distribution in honour of Carl Friedrich Gauss (1777 – 1855), a German mathematician; who derived its equation from a study of errors for repeated measurements of the same quantity. The random variable X having the bell shaped distribution is called a normal random variable. The mathematician equation for the probability distribution of the continuous normal variable depends upon the two parameters μ and σ, its mean and standard deviation. Hence, we denote the probability density function of the normal random variable X by the equation of the normal curve. If X is a normal random variable with mean μ and variance σ^2, then the equation of the normal curve is as follows:

$$f(x,\mu,\sigma)=\frac{1}{\sqrt{2\Pi\sigma}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{for } -\infty < x < \infty$$

Where π = 3.14159 a constant

e = 2.71828

μ is the population mean

σ is the population standard deviation

Whereas the possible values of a discrete random variable can be written as a sequence of isolated values, a continuous random variable is one whose set of possible values is an interval. That is, a continuous random variable is able to take on any value within some interval. For example, such variables as the time it takes to complete a scientific experiment and the weight of an individual are usually considered to be continuous random variables Every continuous random variable X has a curve associated with it. This curve, formally known as a probability density function, can be used to obtain probabilities associated with the random variable. This is accomplished as follows. Consider any two points a and b, where a is less than b.

The probability that X assumes a value that lies between a and b is equal to the area under the curve between a and b. That is,

P{a ≤ X ≤ b} = area under curve between a and b

Since X must assume some value, it follows that the total area under the density curve must equal 1. Also, since the area under the graph of the probability density function between points a and b is the same regardless of whether the endpoints a and b are themselves included, we see that

P{a ≤ X ≤ b} = P{a < X < b}

The normal distribution is a continuous distribution, and is the single most important of all the
distributions discussed in this text. It is widely used and even more widely abused. Its graph is bell shaped, and we see the bell curve in many disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the
normal distribution to help determine your grade. Most IQ scores are normally distributed, as are
many standardized test scores. In this chapter, we will study the normal distribution, the standard
normal distribution, and applications associated with them.

The normal distribution has two parameters (two numerical descriptive measures of the population),
the mean, μ, and the standard deviation, σ. If *X* is a quantity to be measured that has a normal
distribution with mean μ and standard deviation σ, we designate this by writing $X \sim N(,)$. The
probability density function is a rather complicated function. **Do not memorize it**. It is not necessary
for calculations and is included only for completeness:

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

The Standard Normal Distribution

In the normal equation, because π and e are mathematical constants, the probabilities of the random variable X are independent only by μ and σ. Every time a particular combination of μ and σ is specified, a different normal probability distribution is generated. Unfortunately, the normal equation stated above is computationally tedious to use in calculating probabilities. To avoid such computations, a set of tables that provide the desired probabilities are provided. The tables are mostly used in calculating normal probabilities. However, because an infinite number of combinations of the parameters μ and σ exist, an infinite number such tables would be required. By standardizing the data, only one table is needed. By the use of the transformation formula given below, any normal random variable X can be converted to a standardized normal random variable Z. The standard normal score Z is equal to the difference between X and the population mean μ , divided by the standard deviation σ which shown below:

$$f(Z) = \frac{1}{\sqrt{2\Pi}} e^{-\frac{z^2}{2}}$$

The standard normal distribution is a normal distribution of standardized values called ***z*-scores**. *z*-score is measured in units of the standard deviation. For example, if

the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean, since

$$x = \mu + z\sigma = 5 + 3*2 = 11$$

So the *z*-score corresponding to $x = 11$ is $z = 3$. If the value x comes from a normal distribution

with mean μ and standard deviation, , then the transformation

$z = \frac{x - \mu}{\sigma}$ Produces the distribution $Z \sim N(0, 1)$. That is, the mean for the standard normal distribution is zero, and the standard deviation is one.

Properties of the Normal Curve

The properties of the normal curve are as follows:

(1) The mean, median and mode coincide on the horizontal axis where the curve is a maximum at X = μ

(2) The curve is symmetric about a vertical axis through the mean μ

(3) The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean μ

(4) The total area under the curve and above the horizontal axis is equal to 1.

Note:

By the Central Limit Theorem, the normal distribution could be related to some discrete distributions; for example, we have normal approximation to the binomial and normal approximation to the Poisson.

Illustration

Example 1

The heights of poles produced by a manufacturing plant are normally distributed with mean 4.5m and standard deviation 0.5m. Find the probability that a pole selected at random from the plant will have a height;

(a) More than 6.0m

(b) Less than 4.3m

(c) Between 3.4m and 4.9m

Solution

In this particular problem, we have the following;

$\mu = 4.5$

$\sigma = 0.5$

$$Z = \frac{X-\mu}{\sigma}$$

(a) $X = 6.0$

$$Z = \frac{X-\mu}{\sigma} = \frac{6.0-4.5}{0.5} = 3.0$$

$$P(X > 6.0) = P(Z > 3.0) = \Phi(3.0) = 0.00135$$

(b) $X = 4.3$

$$Z = \frac{X-\mu}{\sigma} = \frac{4.3-4.5}{0.5} = -0.4$$

$$P(X < 4.3) = P(Z < -0.4) = 1 - \Phi(-0.4) = 1 - 0.65542 = 0.34458$$

(c) $X_1 = 3.4$

$X_2 = 4.9$

$$Z_1 = \frac{X_1-\mu}{\sigma} = \frac{3.4-4.5}{0.5} = -2.2$$

$$Z_2 = \frac{X_2-\mu}{\sigma} = \frac{4.9-4.5}{0.5} = 0.8$$

$$P(3.4 < X < 4.9) = P(-2.2 < Z < 0.8)$$

$$= \Phi(-2.2) - \Phi(0.8) = 0.98610 - 0.21186 = 0.77424$$

Example 2

The waist measurement of girls of a particular age is normally distributed with mean 66cm and standard deviation 5cm. If six girls of such age are selected at random; find the probability that at least one of them will have a waist measurement of less than 64cm.

Solution

This problem is a combination of normal and binomial distributions since the probability of success p in the binomial distributions will be determined through the normal distribution.

$\mu = 66$

$\sigma = 5$

$X = 64$

$$Z = \frac{X-\mu}{\sigma} = \frac{64-66}{5} = -0.4$$

$$P(X < 64) = P(Z < -0.4) = 1 - \Phi(-0.4) = 1 - 0.65542 = 0.34458$$

The binomial parameters are obtained as follows;

$n = 6$

$p = 0.34$

$q = 0.66$

$$P(X = x) = \binom{n}{x} P^x q^{n-x}$$

$$P(X \geq 1) = 1 - P(X = 0)$$

$$P(X = 0) = \binom{6}{0} (0.34)^0 (0.66)^6 = 0.0826$$

$$P(X \geq 1) = 1 - 0.0826 = 0.9173$$

Example 3

The marks obtained by a large number of students who sat for an examination is normally distributed. If 14% of the students obtained less than 30 marks and 26% obtained more than 50 marks; find the mean and variance of the mark distribution.

Solution

Let mean = μ

Standard deviation = σ

Since 14% of the students obtained less than 30 marks, we have;

$$P(X < 30) = P(Z < \frac{30-\mu}{\sigma}) = 0.14$$

For the convenience of our table, we now change to greater than sign;

$$P(Z > \frac{30-\mu}{\sigma}) = 1 - 0.14 = 0.86$$

$$\frac{30-\mu}{\sigma} = \Phi^{-1}(0.86) = -1.08$$

$\mu - 1.08\sigma = 30$ ....(1.1)

$$P(Z > \frac{50-\mu}{\sigma}) = 0.26$$

$$\frac{50-\mu}{\sigma} = \Phi^{-1}(0.26) = 0.65$$

$\mu + 0.65\sigma = 50$ ....(1.2)

Solving (1.1) and (1.2) simultaneously we have;

The mean $\mu = 42.5$

The standard deviation $\sigma = 11.6$

The variance $\sigma^2 = (11.6)^2 = 134.6$

Summary

In this unit you have learnt:

- How to obtain values for Normal distribution and that the value obtained must be between 0 and 1
- How to obtain the expected value and variance.

Self Assessment Questions

(1) Describe the Normal distribution

(2) The uses of normal distribution

(3) Describe the properties of normal distribution

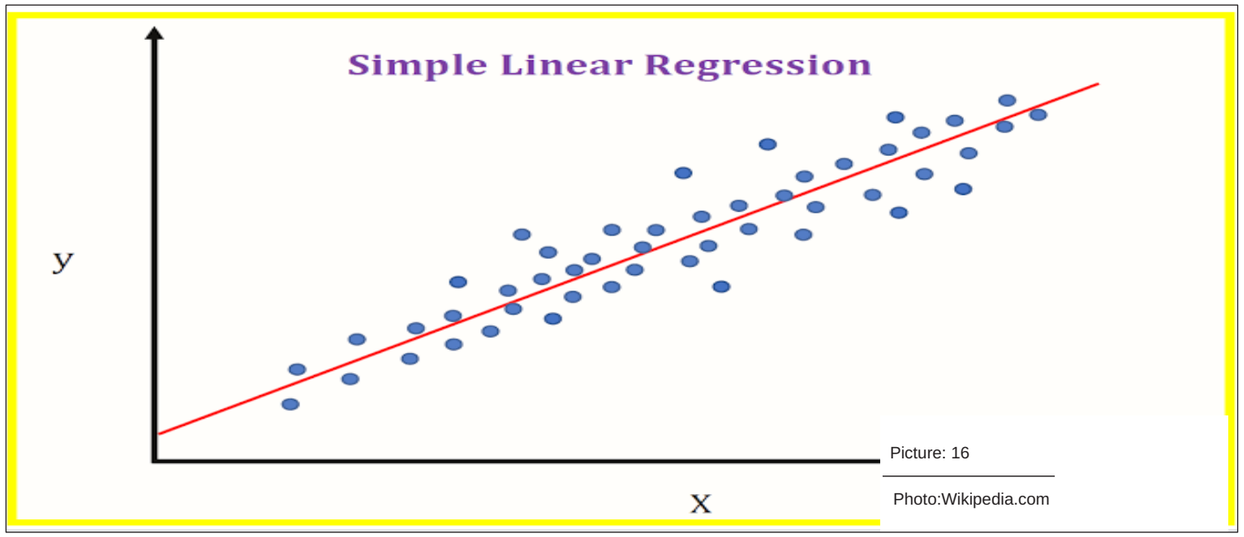
(4) Compute normal distribution

Tutor Marked Assessment

- A certain car comes in six colours, two engine sizes and two transmission types. A dealer has 20 of these cars in stock. A customer wants a red car, manual transmission and big engine. Find the probability that the dealer has such car in stock.
- The ages of students in a class is normally distributed with mean 20 years and variance 9. Find the probability that the age of a randomly selected student will lie between 26 and 29 years.
- The final exam scores in a large statistics class were normally distributed with a mean of 63 and a standard deviation of five. Find the probability that a randomly selected student scored more than 65 on the exam.
- The final exam scores in a large statistics class were normally distributed with a mean of 63 and a standard deviation of five. Find the probability that a randomly selected student scored less than 85 on the exam.
- The final exam scores in a large statistics class were normally distributed with a mean of 63 and a standard deviation of five.
- Find the probability that a randomly selected student scored more than 65 on the exam.
- Find the probability that a randomly selected student scored less than 85 on the exam.
- Find the 90th percentile. That is, find the score that separates the lower 90% of scores from the top 10%.

References

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and
- Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.
- Gupta (2014). Introduction to Fundamental Statistics.
- Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.
- Introductory Statistics 3rd Edition by Sheldom M. Ross



Picture: 16

Photo:Wikipedia.com

UNIT 2

Linear Regression

Introduction

We are often interested in trying to determine the relationship between a pair of variables. For instance, how does the amount of money spent in advertising a new product relate to the first month's sales figures for that product? Or how does the amount of catalyst employed in a scientific experiment relate to the yield of that experiment? Or how does the height of a father relate to that of his son?

In many situations the values of the variables are not determined simultaneously in time; rather, one of the variables will be set at some value, and this will, in turn, affect the value of the second variable. For instance, the advertising budget would be set before the sales figures are determined, and the amount of catalyst to be used would be set before the resulting yield could be determined. The variable whose value is determined first is called the input or independent variable and the other is called the response or dependent variable.

Learning Outcomes

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

1. Know the meaning of linear regression.
2. Express linear regression.
3. Express the difference between linear and multiple regression.

Main Content

8 mins

Here, we are often interested in trying to determine the relationship between a pair of variables. For instance, how does the amount of money spent in advertising a new product relate to the first month's sales figures for that product? Or how does the amount of catalyst employed in a scientific experiment relate to the yield of that experiment? Or how does the height of a father relate to that of his son?

In many situations the values of the variables are not determined simultaneously in time; rather, one of the variables will be set at some value, and this will, in turn, affect the value of the second variable. For instance, the advertising budget would be set before the sales figures are determined, and the amount of catalyst to be used would be set before the resulting yield could be determined. The variable whose value is determined first is called the input or independent variable and the other is called the response or dependent variable.

Suppose that the value of the independent variable is set to equal x. Let Y denote the resulting value of the dependent variable. The simplest type of relationship between this pair of variables is a straight-line, or linear, relation of the form

$$Y = \alpha + \beta x$$

This model, however, supposes that (once the values of the parameters α and β are determined) it would be possible to predict exactly the response for any value of the input variable. In practice, however, such precision is almost never attainable, and the most that one can expect is that the preceding equation is valid subject to random error.

Simple Linear Regression Model

Let us consider a pair of variables, one of which is called the input variable and the other the response variable. Suppose that for a specified value x of the input variable the

$$Y = \alpha + \beta x + e$$

The quantities α and β are parameters. The variable e, called the random error, is assumed to be a random variable having mean 0.

Regression is the relationship between the response variable Y and the input variable x specified in the preceding equation is called a simple linear regression. The simple linear regression relationship can also be expressed by stating that for any value x of the input variable, the response variable Y is a random variable with mean given by

$$E[Y] = \alpha + \beta x$$

Thus a simple linear regression model supposes a straight-line relationship between the mean value of the response and the value of the input variable Parameters α and β will almost always be unknown and will have to be estimated from data.

To see if a simple linear regression might be a reasonable model for the relationship between a pair of variables, one should first collect and then plot data on the paired values of the variables. For instance, suppose there is available a set of data pairs (xi, yi), i = 1, . . . , n, meaning that when the input variable was set to equal xi, the observed value of the response variable was yi. These points should then be plotted to see if, subject to random error, a straight-line relationship between x and y appears to be a reasonable assumption. The resulting plot is called a scatter diagram.

Estimating Regression Parameters

Suppose that the responses Yi corresponding to the input values xi, i = 1, . . . , n, are to be observed and used to estimate the parameters α and β in a simple linear regression model

$$Y = \alpha + \beta x + e$$

To determine estimators of α and β, we reason as follows: If A and B were the respective estimators of α and β, then the estimator of the response corresponding to the input value xi would be A + Bxi. Since the actual response is Yi, it follows that the difference between the actual response and its estimated value is given by

$$€i \equiv Yi - (A + Bxi)$$

That is, i represent the error that would result from using estimators A and B to predict the response at input value xi. Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be influenced by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by

an initial amount plus an hourly fee. The type of data described in this chapter is **bivariate** data –

the prefix "bi" indicating there are two variables. Although we will only study models involving two variables, in many real-world situations statisticians use **multivariate** data, meaning many variables.

In this study, we will be studying the simplest form of regression, "linear regression" with one

independent variable. More specifically, given sample data measured on two variables, *x* and *y*, we wish to find a linear equation that best fits the observed sample data. We will also develop statistical measures and tests that measure how strong the linear relationship is between the variables.

We start by reviewing the basic facts about linear equations. A **linear equation** is an equation of

the form:

$$y = a + bx$$

Where *a* and *b* are constant numbers. This is called a linear equation, because the graph of the

equation is a straight line. The variable *x* is called the **independent variable**, and *y* is the **dependent variable.** That is, as written, *y* depends on *x*. Given any value, we can substitute it for the independent variable *x* to obtain the corresponding value for the dependent variable.

Slope and Intercept of a Linear Equation

Given a linear equation y = a + bx, we call b the slope, and a is called the y-intercept. From basic algebra, recall that the slope is a number that describes the steepness of the line, and the y-intercept is the y coordinate of the point (0, a) where the line crosses the y-axis. The sign of the slope b

determines whether the line slopes upward or downward.

(a) If b > 0, the line slopes upward to the right.

(b) If b = 0, the line is horizontal.

(c) If b < 0, the line slopes downward to the right.

More specifically, when the slope b is positive, an increase in x results in an increase in y. And if the slope b is negative, then an increase in x results in a decrease in y.

Scatter Plots

Before we begin our discussion of linear regression and correlation, we need to a way display the relation between two variables *x* and *y*. The most commonly used graph is a **scatter plot** (also called a scatter diagram). To make a scatterplot, we just plot the points, for each data pair using the *x* value as the *x*-coordinate and the *y*-value as the *y*-coordinate. For small examples, this is easy to do by hand; but most statistical software packages will make very nice scatterplots,

The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to fit a line to the points in the scatter plot. This line can be calculated through a process called **linear regression**, which we will learn in the next section. We will also develop a numerical measure that provides a more objective measure of how well the line fits the data. And we will only use the fitted line when we have determined that there is a statistically significant relationship between the variables *x* and *y*.

Illustration

Example 1

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret these using complete sentences.

Solution

The independent variable x is the number of hours Svetlana tutors each session. The dependent variable y is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each additional hour she tutors, Svetlana earns an additional $15.

Example 2

The following sample data was obtained is a study of the relationship between the number of years that foreign students study English language in Nigerian Universities and the scores that they received in a proficiency test (%) in that language:

Number of years spent and scores

Years spent(X)	3	4	4	2	5	3	4	5	3	2
Test Scores(Y)	57	78	72	58	89	63	73	84	75	48

You are expected to use the method of least squares to obtain a linear regression equation for predicting the students' score given the number of years of study. That is an equation of the form; $\hat{Y} = a + bX_i$

Solution

The sums needed for substitution into the normal equations are obtained by performing the calculations shown in the following table:

Number of years(X)	Scores in Test (Y)	X^2	XY
3	57	9	171
4	78	16	312
4	72	16	288
2	58	4	116
5	89	25	445
3	63	9	189
4	73	16	292
5	84	25	420
3	75	9	225
2	48	4	96
35	697	133	2554

Hence, from the table we have:

n = 10

$$\sum_{i=1}^{n} X_i = 35$$

$$\sum_{i=1}^{n} Y_i = 697$$

$$\sum_{i=1}^{n} X_i^2 = 133$$

$$\sum_{i=1}^{n} X_i Y_i = 2554$$

Here we need the following least squares, linear regression equation:

$$b = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{10\times 2554 - 35\times 697}{10\times 133 - (35)^2} = 10.90$$

$$a = \frac{\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i}{n} = \frac{697 - 10.90\times 35}{10} = 31.55$$

Hence the required least square linear regression equation is as follows:

$\hat{Y} = a + bX_i$

Using the least square equation we can estimate the test scores of a student who spent six years of study as follows:

In this case, X = 6 and we substitute into the equation:

$\hat{Y}_i = 31.55 + 10.90 \times 6 \approx 97$

In other words, a student who spent six years of study are expected to score an average of 97% in the test.

Example 3

Consider the following data, where x denotes the respective number of branches that 10 different banks have in some metropolitan area and y denotes the corresponding share of the total deposits held by the banks:

X	198	186	116	89	120	109	28	58	34	31
Y	22.7	16.6	15.9	12.5	10.2	6.8	6.8	4.0	2.7	2.8

Fit the least square model

Solution

$$b = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{10 \times 2{,}286555 - 690 \times 33{,}140}{10 \times 47{,}816 - (690)^2} = -0.2959$$

$$a = \frac{\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i}{n} = \frac{101 - (-0.2959) \times 690}{10} = 17.8100$$

Hence the required least square linear regression equation is as follows:

$\hat{Y}_i = 17.8100 - 0.2959X_i$

Summary

In this unit you have learnt:

- How to obtain values for Regression Analysis and
- How to obtained the least square by fitting the model.

Self Assessment Questions

(1) Know the meaning of linear regression.

(2) Express linear regression.

(3) Express the difference between linear and multiple regression.

Tutor Marked Assessment

- Consider the following list of data value

X	4	2	10	5	8
Y	8	12	4	10	2

Fit the least squares regression.

- The following ten observation on variable X and Y

X	2.5	5	10	15	17.5	20	25	30	35	40
Y	63	58	55	61	62	37	38	45	46	19

use the method of least squares to obtain a linear regression equation

- The following sample data was obtained is a study of the relationship between the number of years that foreign students English language in Nigerian universities and the scores that they received in a proficiency test in that language:

X	3	4	4	2	5	3	4	5	3	2
Y	57	78	72	58	89	63	73	84	75	48

- You are expected to use the method of least squares to obtain a linear equation for predicting the students' score given the number of years of study.

- The paired data below consist of the temperatures on randomly chosen days and the amount a certain kind of plant grew (in millimeters):

Temp	62	76	50	51	71	46	51	44	79
Growth	36	39	50	13	33	33	17	6	16

References

Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and

Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.

Gupta (2014). Introduction to Fundamental Statistics.

Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.

Introductory Statistics 3rd Edition by Sheldom M. Ross

Formula

for X on Y

$$X = a + bY$$

The normal equations are

$$\sum X = na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

Regression coefficient

$$b_{XY} = \frac{\sum XY}{\sum Y^2}$$

Regression equation

$$x = \bar{X} + b_{XY}(Y - \bar{Y})$$

for Y on X

$$Y = a + bX$$

The normal equations are

$$\sum Y = na + b\sum x$$

$$\sum XY = a\sum x + b\sum x^2$$

Regression coefficient

$$b_{YX} = \frac{\sum XY}{\sum x^2}$$

Regression equation

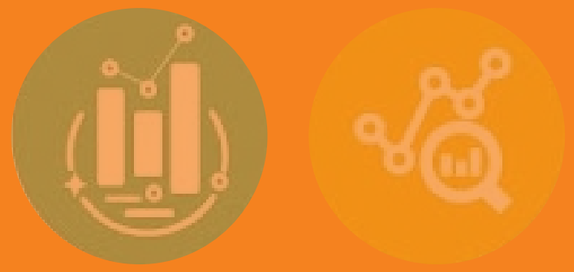
$$Y = \bar{Y} + b_{YX}(X - \bar{X})$$

Picture: 17

Photo:Wikipedia.com

Module 6

Analysis Of Variance For Simple Linear Regression



Picture: 18

Photo:Unsplash.com

UNIT 1

Analysis of Variance for Simple Linear Regression

Introduction

We present a general approach, called the analysis of variance (ANOVA), for making inferences about the mean values of a variety of random variables. In one-factor ANOVA, the mean of a variable depends on only a single factor, namely, the sample to which it belongs. In two-factor ANOVA, the random variables are thought of as being arrayed in a rectangular arrangement, and the mean of a variable depends on both its row and its column factor. We show how to test the hypothesis that the mean of a random variable does not depend on which row the random variable is in, as well as the analogous hypothesis that the mean does not depend on which column it is in.

Learning Outcomes

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

1. Test for adequacy of regression equation
2. Test for the significance of the parameters in the model.
3. Use Analysis of variance in Regression.

Main Content

7 mins

You should be aware that many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

One use of the F distribution is testing two variances. It is often desirable to compare two variances rather than two averages. For instance, college administrators would like two college professors grading exams to have the same variation in their grading. In order for a lid to fit a container, the variation in the lid and the container should be the same. A supermarket might be interested in the variability of check-out times for two checkers.

In order to perform an F test of two variances, it is important that the following are true:

(1) The populations from which the two samples are drawn are normally distributed.

(2) The two populations are independent of each other.

Unlike most other tests in this book, the F test for equality of two variances is very sensitive to deviations from normality. If the two distributions are not normal, the test can give higher p-values than it should, or lower ones, in ways that are unpredictable. Many texts suggest that students not use this test at all, but in the interest of completeness we include it here. We will again be using the 5 steps of hypothesis testing.

Suppose we sample randomly from two independent normal populations. Let 21s and 22s be the

Population variances and 21S and 22Sbe the sample variances. Let the sample sizes be 1n and 2n.

Since we are interested in comparing the two sample variances, we use the F ratio:

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$$

F has the distribution similar to chi-square distribution but dependent on two degrees of freedom.

F Distribution Has The Following Characteristics

(1) All F values are greater than or equal to 0

(2) There is a different F curve for each pair of degrees of freedom n1 – 1, n2 – 1

(3) The curve is non symmetrical and skewed to the right

(4) There is 100% under the curve

(5) The notation is F ~ F(n1 – 1, n2 – 1) where n1 – 1, are the degrees of freedom for the numerator and n2 – 1 are the degrees of freedom for the denominator.

(6) $\mu = \frac{d.f.N}{d.f.D.-1}$

One of the most popular tests of significance for regression parameters is the analysis of variance (ANOVA). In the ANOVA, we are interested in analyzing the variation in Y into its component parts; one part due to relationship with X and the other parts due to error. The general form of the ANOVA table for a simple regression is shown

Source of Variation	Sums of Squares	Degree of freedom	Means squares	F
Regression	SSR = $bS_{xy} = b^2S_{xx}$	1	MSR = $SSR/1$	F = SSR/MSE
Error	SSE = $S_{yy} - bS_{xy}$	n – 2	MSE= $SSE/n-2$	
Total	SST = S_{yy}	n - 1		

Where;

SST is the total sum of squares

SSR is the sum of square due to regression

SSE is the error sum of squares

MSR is the mean square due to regression

MSE is the mean square error.

Where;

$$S_{xx} = \sum_{i=1}^{n}(X-\bar{X})^2 = \sum_{i=1}^{n}X^2 - \frac{\left(\sum_{i=1}^{n}X\right)^2}{n} = \sum_{i=1}^{n}X^2 - n(\bar{X})^2$$

$$S_{yy} = \sum_{i=1}^{n}(Y-\bar{Y})^2 = \sum_{i=1}^{n}Y^2 - \frac{\left(\sum_{i=1}^{n}Y\right)^2}{n} = \sum_{i=1}^{n}Y^2 - n(\bar{Y})^2$$

$$S_{xy} = \sum_{i=1}^{n}(X-\bar{X})(Y-\bar{Y}) = \sum_{i=1}^{n} XY - \frac{\left(\sum_{i=1}^{n} X\right)\left(\sum_{i=1}^{n} Y\right)}{n} = \sum_{i=1}^{n} XY - n(\bar{X}\bar{Y})$$

The ANOVA could be used, especially in multiple regression analysis, to test for the significance of all the regression coefficients. In simple regression the ANOVA is equivalent to t –test since there is only one coefficient as $F = t^2$.

The hypothesis for ANOVA is as follows:

Hypothesis

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ (All the coefficients are not significant)

Vs

$H_1 : \beta_1 \neq \beta_2 \neq ... \neq \beta_k \neq 0$ (At least one of the coefficients is significant)

Test Statistic

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

Decision Criterion

Reject H_0 if $F \geq F_{\alpha,1,n-2}$ for α level of significance.

s$t = \frac{\alpha - \alpha_0}{SE(\alpha)}$ **st in Regression:**

The student's t –test can be used to individually test hypothesis about either the slope or the intercept of a regression equation. For the slope parameter 'b', the procedure for the test of hypothesis is as follows:

Hypothesis

$H_0 : \beta = b_0$ (the slope coefficients is equal to a specified constant b_0)

$H_1 : \beta \neq b_0$ (the slope coefficients is not equal to a specified constant b_0)

Test Statistic

$$b = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

b_0 is the hypothesized value in H_0 but if $b_0 = 0$ in H_0 then the test reduce to more test of significance as follows:

$H_0 : \beta = 0$ (The slope coefficients is not significance)

$H_1 : \beta \neq 0$ (The slope coefficients is significance)

Test Statistic

$$t = \frac{b}{SE(b)}$$

$$SE(b) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Where;

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{S_{yy} - b^2 S_{xx}}{n-2}$$

$$S_{xx} = \sum_{i=1}^{n}(X-\bar{X})^2 = \sum_{i=1}^{n} X^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n} = \sum_{i=1}^{n} X^2 - n(\bar{X})^2$$

$$S_{yy} = \sum_{i=1}^{n}(Y-\bar{Y})^2 = \sum_{i=1}^{n} Y^2 - \frac{\left(\sum_{i=1}^{n} Y\right)^2}{n} = \sum_{i=1}^{n} Y^2 - n(\bar{Y})^2$$

Decision Criterion

Reject H_0 if $t \geq t_{\frac{\alpha}{2},n-2}$ for α level of significance.

For the intercept parameter 'α' the procedure for the test hypothesis is as follows:

Hypothesis

$H_0 : \alpha = \alpha_0$ (The intercept parameter is equal to a specified constant α_0)

Vs

$H_1 : \alpha \neq \alpha_0$ (The intercept parameter is not equal to a specified constant α_0)

Test Statistic

$$t = \frac{\alpha - \alpha_0}{SE(\alpha)}$$

where

$$\alpha = \frac{\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i}{n}$$

0a is the hypothesized value in H_0 but if $\alpha_0 = 0$ in H_0 then the test reduce to mere test of significance as follows:

$H_0 : \alpha = \alpha_0$ (The intercept parameter is not significant)

$H_1 : \alpha \neq \alpha_0$ (The intercept parameter is significant)

$$t = \frac{\alpha}{SE(\alpha)}$$

Where

$$SE(\alpha) = \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$SE(\alpha) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]}$$

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{S_{yy} - b^2 S_{xx}}{n-2}$$

$$S_{xx} = \sum_{i=1}^{n}(X - \bar{X})^2 = \sum_{i=1}^{n} X^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n} = \sum_{i=1}^{n} X^2 - n(\bar{X})^2$$

$$S_{yy} = \sum_{i=1}^{n}(Y - \bar{Y})^2 = \sum_{i=1}^{n} Y^2 - \frac{\left(\sum_{i=1}^{n} Y\right)^2}{n} = \sum_{i=1}^{n} Y^2 - n(\bar{Y})^2$$

Decision Criterion

Reject H_0 if $t \geq t_{\frac{\alpha}{2},n-2}$ for α level of significance.

Illustration

Example 1

The following sample data was obtained is a study of the relationship between the number of years that foreign students English language in Nigerian universities and the scores that they received in a proficiency test in that language:

X	3	4	4	2	5	3	4	5	3	2
Y	57	78	72	58	89	63	73	84	75	48

You are expected to use the method of least squares to obtain a linear equation for predicting the students' score given the number of years of study. Hence, conduct ANOVA test for the parameter of the regression model.

Solution

$n = 10$

$$\sum_{i=1}^{n} X_i = 35$$

$$\sum_{i=1}^{n} Y_i = 697$$

$$\sum_{i=1}^{n} X_i^2 = 133$$

$$\sum_{i=1}^{n} X_i Y_i = 2554$$

Here we need the following least squares, linear regression equation:

$$b = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{10 \times 2554 - 35 \times 697}{10 \times 133 - (35)^2} = 10.90$$

$$a = \frac{\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i}{n} = \frac{697 - 10.90 \times 35}{10} = 31.55$$

Hence the required least square linear regression equation is as follows:

$$\hat{Y}_i = 31.55 + 10.90X_i$$

The ANOVA test is conducted as follows

Hypothesis

$H_0 : \alpha = \alpha_0$ (the slope coefficients is not significant)

Vs

$H_1 : \alpha \neq \alpha_0$ (the slope coefficients is significant)

Level of significance

α = 0.05

Test Statistic

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

Decision Criterion

Reject H_0 if $F \geq F_{0.05,1,8} = 5.32$ at the 5% level of significance.

Computation

$$SST = S_{yy} = \sum_{i=1}^{n}(Y - \bar{Y})^2 = \sum_{i=1}^{n} Y^2 - \frac{\left(\sum_{i=1}^{n} Y\right)^2}{n} = \sum_{i=1}^{n} Y^2 - n(\bar{Y})^2 = 50085 - \frac{(697)^2}{10} = 1504.1$$

$$S_{xx} = \sum_{i=1}^{n}(X - \bar{X})^2 = \sum_{i=1}^{n} X^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n} = \sum_{i=1}^{n} X^2 - n(\bar{X})^2 = 133 - \frac{(35)^2}{10} = 10.5$$

$$SSR = b^2 S_{xx} = (10.90)^2 \times 10.5 = 1247.51$$

and

SSE = SST – SSR = 15041 – 1247.51 = 256.59

The rest of the computations are shown in the ANOVA table below:

Source of Variation	Sums of squares	Degree of freedom	Mean squares	F
Regression	1247.51	1	1247.51	38.89
Error	256.59	8	32.074	
Total	1504.10	9		

The conclusion follows immediately,

Conclusion

Since F = 38.89 exceeds 5.32, we reject H_0 and conclude that the slope parameter is statistically significant.

Example 2

The following sample data was obtained is a study of the relationship between the number of years that foreign students English language in Nigerian universities and the scores that they received in a proficiency test in that language:

X	3	4	4	2	5	3	4	5	3	2
Y	57	78	72	58	89	63	73	84	75	48

You are expected to use the method of least squares to obtain a linear equation for predicting the students' score given the number of years of study. Hence, conduct t - test for the slope and intercept parameters of the regression model.

Solution

n = 10

$$\sum_{i=1}^{n} X_i = 35$$

$$\sum_{i=1}^{n} Y_i = 697$$

$$\sum_{i=1}^{n} X_i^2 = 133$$

$$\sum_{i=1}^{n} X_i Y_i = 2554$$

Here we need the following least squares, linear regression equation:

$$b = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{10 \times 2554 - 35 \times 697}{10 \times 133 - (35)^2} = 10.90$$

$$a = \frac{\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i}{n} = \frac{697 - 10.90 \times 35}{10} = 31.55$$

Hence the required least square linear regression equation is as follows:

$$\hat{Y}_i = 31.55 + 10.90X_i$$

The t – test for the slope parameter is conducted as follows

Hypothesis

$H_0 : \alpha = \alpha_0$ (the slope coefficients is not significant)

Vs

$H_1 : \alpha \neq \alpha_0$ (the slope coefficients is significant)

Level of significance

α = 0.05

Test Statistic

$$t = \frac{b}{SE(b)}$$

Decision Criterion

Reject H_0 if $t \geq t_{0.025,8} = 2.306$ at the 5% level of significance.

Computation

$$SST = S_{yy} = \sum_{i=1}^{n}(Y - \bar{Y})^2 = \sum_{i=1}^{n} Y^2 - \frac{\left(\sum_{i=1}^{n} Y\right)^2}{n} = \sum_{i=1}^{n} Y^2 - n(\bar{Y})^2 = 50085 - \frac{(697)^2}{10} = 1504.1$$

$$S_{xx} = \sum_{i=1}^{n}(X - \bar{X})^2 = \sum_{i=1}^{n} X^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n} = \sum_{i=1}^{n} X^2 - n(\bar{X})^2 = 133 - \frac{(35)^2}{10} = 10.5$$

$$SSR = b^2 S_{xx} = (10.90)^2 \times 10.5 = 1247.51$$

and

SSE = SST – SSR = 15041 – 1247.51 = 256.59

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{256.59}{8} = 32.074$$

$$SE(b) = \sqrt{\frac{\hat{\sigma}^2}{n\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \sqrt{\frac{32.074}{10.5}} = 1.75$$

$$t = \frac{b}{SE(b)} = \frac{10.90}{1.75} = 6.23$$

Conclusion

Since t = 6.23 exceeds 2.306, we reject the null hypothesis and conclude that the slope parameter is statistically significant.

The t –test for the intercept parameter is conducted as follows:

Hypothesis

$H_0 : \alpha = \alpha_0$ (the intercept coefficient is not significant)

Vs

$H_1 : \alpha \neq \alpha_0$ (the intercept coefficient is significant)

Level of significance

α = 0.05

Test Statistic

$$t = \frac{\alpha}{SE(\alpha)}$$

Decision Criterion

Reject H_0 if $t \geq t_{0.025,8} = 2.306$ at the 5% level of significance.

Computation

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{256.59}{8} = 32.074$$

$$SE(\alpha) = \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \sqrt{\frac{32.074 \times 133}{10.5 \times 10}} = 6.374$$

$$t = \frac{\alpha}{SE(\alpha)} = \frac{31.55}{6.374} = 4.95$$

Conclusion

Since t = 4.95 exceeds 2.306, we reject the null hypothesis and conclude that the intercept parameter is statistically significant.

Summary

In this unit you have learnt:

- How to obtain values for Analysis of variance for regression and
- How to obtained the value of parameters

Self Assessment Questions

(1) Test for adequacy of regression equation

(2) Test for the significance of the parameters in the model.

(3) Use Analysis of variance in Regression.

Tutor Marked Assessment

- An investigator for a consumer cooperative organized a study of the mileages obtainable from three different brands of gasoline. Using 15 identical motors set to run at the same speed, the investigator randomly assigned each brand of gasoline to 5 of the motors. Each of the motors was then run on 10 gallons of gasoline, with the total mileages obtained as follows.

Gas 1	Gas 2	Gas 3
220	244	252
251	235	272
226	232	250
246	242	238
260	225	256

- Test the hypothesis that the average mileage obtained is the same for all three types of gasoline. Use the 5 percent level of significance.

A random sample of size 3 is taken from each of three independent, normally distributed random variables X_1, X_2, X_3, having equal but unknown variances. Test at the 0.05 level of significance, the hypothesis that X_1, X_2, X_3, have equal means.

X_1	94	82	84
X_2	102	94	78
X_3	76	68	70

- A home gardener wishes to determine the effect of different fertilizers on the average number of tomatoes produced by her plants. She grows five tomato plants on each of four separate plots, X_1, X_2, X_3, X_4, and uses a different fertilizer treatment on each plot. The number of tomatoes per plant are indicated in the following table. Test, at the 0.05 level of significance, the hypothesis that plots X_1, X_2, X_3, and X_4 have equal average yields.

X_1	14	10	12	16	17
X_2	9	11	12	8	10
X_3	16	15	14	10	18
X_4	10	11	11	13	8

References

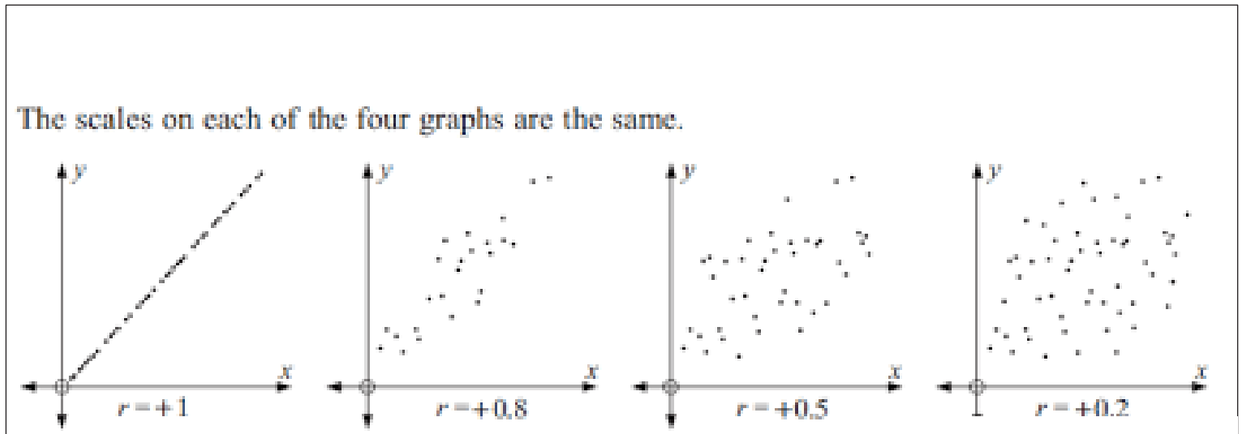
Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and

Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.

Gupta (2014). Introduction to Fundamental Statistics.

Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.

Introductory Statistics 3rd Edition by Sheldom M. Ross



Picture: 19

Photo:Wikepedia.com

UNIT 2

Correlation

Introduction

Consider a set of data pairs (xi, Yi), i = 1,... , n.we defined the sample correlation coefficient of this data set by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

It was noted that r provided a measure of the degree to which high values of x are paired with high values of Y and low values of x with low values of Y. A value of r near +1 indicated that large x values were strongly associated with large Y values and small x values were strongly associated with small Y values, whereas a value near −1 indicated that large x values were strongly associated with small Y values and small x values with large Y values.

Learning Outcomes

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

1. Explain correlation
2. Express how the correlation is obtain from two or more variable.
3. Show the degree of relationship between the variables whether strong or weak

Main Content

7 mins

Let us consider the data set of paired values (x1, y1),(x2, y2),... ,(xn, yn). In this section we will present a statistic, called the sample correlation coefficient that measures the degree to which larger x values go with larger y values and smaller x values go with smaller y values.

We are also interested in determining the strength of the relationship between a pair of variables in which large values of one variable tend to be associated with small values of the other. we observed that higher numbers of years of schooling tend to be associated with lower resting pulse rates and that lower numbers of years of schooling tend to be associated with the higher resting pulse rates. This is an example of a negative correlation.

Correlation may be defined as the measure of the degree and direction of linear relationship existing between two or more variable capable of quantitative measurement. The strength of a relationship or the association between two variables is typically measured by the coefficient of correlation whose value range from -1 for a perfect negative correlation up to +1 for a perfect positive correlation. The correlation coefficient is a useful and versatile statistic, but one must be careful not to confuse correlation with causation. For instance, a high correlation between smoking and lung cancer, one may be tempted to suspect that smoking somehow causes physiological or metabolic changes in the lunges. This is propably true, but in general, a high correlation between X and Y (that is when the absolute correlation coefficient is near 1) does not necessarily imply that X causes Y or that Y causes X. It may simply means that some other factor W or some combinations of factors, influences both X and Y. For example, educational research has established children's shoe size (X) is highly correlated with spelling ability (Y). This is simply because shoe size and spelling performance are both correlated with age (W); older children have bigger feet than younger children and older children spell better.

When only two variables are involved, we speak of simple correlation. On the other hand, when more than two variables are involved, we speak of multiple or partial correlations. This text will only consider simple correlation

Types of Correlation

We can observe the type of correlation between two variables through a graphical display. The graphical display of the two variables in the (x,y) plane is often called the scatter diagram. The three types of correlation are discussed as follows:

(1) Positive correlation
(2) Negative correlation
(3) No correlation

Positive correlation

Two variables are positively correlated if they tend to increase and decrease together in the same direction. If all points lie on a straight line the correlation is said to be perfect positive. The closer the value of the positive correlation coefficient is to +1, thestronger the degree of positive relationship that exists between the two variables. This is depicted in the scatter.

Negative correlation

Two variables are negatively correlated if they tend to increase and decrease together in the opposite direction. If all the points lie on a straight line, the correlation is said to be perfect negative. The closer the value of the negative correlation is to 1, the stronger the degree of negative relationship that exists between the two variables. This is depicted in the scatter.

No correlation

Two variables are uncorrelated if they tend to change with no definite connection to each depicted in the scatter.

Measures of Correlation

Correlation was first investigated graphically by Sir Francis Galton, for this reason, the scatter diagram is often called the Galton graph. In 1896, Karl Pearson proposed a method of assessing correlation by a formula based upon a mathematical study of the regression lines.

Scatter Diagram

The scatter diagram is otherwise called the Galton graph in honour of its originator, Sir Francis Galton, an English biometrician. The type of correlation between two variables can be determined by the direct observation of the scatter diagram of the two variables. The scatter diagram is a qualitative picture on how well a given line or curve describes the relationship between variables. If the points on the scatter diagram lie close to the line then the correlation is strong otherwise weak. This method does not give a quantitative measurement of the correlation between variable. It is rather based on individual observation of the scatter plot. If X and Y represent the two variable under consideration, a scatter diagram shows the location of points (x,y) on a rectangular coordinate system. If all points in the scatter diagram seem to lie near a line, the observation is called linear. In such case, a linear equation is appropriate for purposes of regression or estimation. In order to draw inferences about the strength of relationship between variables, a quantitative measure is necessary rather that a scatter plot.

Product – Moment Correlation Coefficient

It will interest you to note that Product-Moment Correlation Coefficient is otherwise called the Karl Pearson's product – moment correlation coefficient in honour of Karl Pearson, an English statistician who derived its formula in 1896. The product moment correlation coefficient between two variables x and y, denoted by r for population and r for sample, is given by the formula.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}}$$

where

$$S_{xy} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$ is the sample covariance between x and y

$$S_{xx} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$$ is the sample variance of x

$$S_{yy} = \frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}$$ is the sample variance of y

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$$ is the sample mean of x

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}Y_i$$ is the sample mean of y

Note

By analogy, the deviation from the mean $\sum_{i=1}^{n}(X_i-\bar{X})$ and $\sum_{i=1}^{n}(Y_i-\bar{Y})$ are often called the first moments about the mean, while $\sum_{i=1}^{n}(X_i-\bar{X})^2$ and $\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ are called the second moments. A convenient name for $\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$ is therefore product – moment. From this analogy that is the reason why r is known as the product – moment correlation coefficient.

$$r = \frac{n\sum_{i=1}^{n}X_iY_i - \sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{\sqrt{\left(n\sum_{i=1}^{n}X^2-\left(\sum_{i=1}^{n}X\right)^2\right)\left(n\sum_{i=1}^{n}Y^2-\left(\sum_{i=1}^{n}Y\right)^2\right)}}$$

For $-1 \le r \le +1$

That is, the above correlation coefficient lies between -1 and +1 inclusive. The closer the value of r is to 1± the stronger the degree of the relationship that exists between the two variables.

Coefficient of Determination

Coefficient of determination denoted by r^2 is the square of the product – moment correlation coefficient. In regression analysis, the coefficient of determination is percentage of variation in the dependent variable that was due to the independent variable. The coefficient of determination is thus given by the formula:

$$r^2 = \frac{n\sum_{i=1}^{n}X_iY_i - \sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{\left[n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right]\left[n\sum_{i=1}^{n}Y^2-\left(\sum_{i=1}^{n}Y_i\right)^2\right]}$$

For $-1 \leq r \leq +1$

That is, the coefficient of determination lies between 0 and +1 inclusive. The closer the value of r^2 is to +1 the stronger the degree of relationship that exists between the two variables. However, the coefficient of determination is better interpreted during regression analysis where the dependent and independent variables are clearly defined.

Caution in Interpreting Correlation Coefficient

Interpreting analyses of correlation between two variables became an important question in statistics during the increasingly widespread use of correlation methods in the early 19th century. Karl Pearson knew that a large correlation between two variables could be due to their correlation with a third variable. This phenomenon was not recognized until 1926 when George Udny Yule proved it through an example by getting the coefficient of correlation between time series. We have identified three of the most common errors made in interpreting results involving correlation coefficients. These errors are as follows:

(1) We must be very careful to avoid concluding that a significant linear correlation less is a between two variables is a proof that there is a cause – and – effect relationship between the X and Y variables. For instances, the statistical correlation between smoking and cancer is not enough proof to say that smoking causes cancer. The correlation techniques can be used only to establish a linear relationship. The techniques alone cannot establish the existence or absence of any inherent cause – and – effect relationship between the variables.

(2) Another source of potential error arises with data based on rates or averages. When we use rates or averages of data, we suppress the variation among the individuals or items, and this may easily lead to an inflated correlation coefficient. One study produced a linear correlation coefficient of 0.4 for paired data relating income and education among individuals, but the correlation coefficient between 0.7 when regional averages were used.

(3) A third source of error involves the property of linearity. The conclusion that there is no significant linear correlation does not mean that X and Y are not related in any way. A data may produce zero correlation coefficient, an indication of no linear relationship between the two variables. However, the same data may produce a strong non – linear relationship. The use of scatter diagram can always assist in overcoming this error.

A remedial measure to the problems of bivariate correlations is the use of partial correlation. The partial correlation between two variables is defined as correlation of two variables while controlling for a third or more other variables. A measure of partial correlation between variables X and Y has a third variable Z as a measure of the direct relation between X and Y that does not take into accounts the consequences of the relations of these two variables with Z.

Illustration

Example

Compute and interpret the coefficient of correlation and the coefficient of determination for the sample data below.

X	20.4	19.7	21.8	20.1	20.7
Y	9.2	8.9	11.4	9.4	10.3

Solution

The computation are summarized in the following table;

X_i	Y_i	X_{i}^2	Y_i^2	X_iY_i
20.4	9.2	416.16	84.64	187.68
19.7	8.9	388.09	79.21	175.33
21.8	11.4	475.24	129.96	248.52
20.1	9.4	404.01	88.36	188.94
20.7	10.3	428.49	106.09	213.21
$\sum_{i=1}^{n} X_i$	$\sum_{i=1}^{n} Y_i$	$\sum_{i=1}^{n} X_i^2$	$\sum_{i=1}^{n} Y_i^2$	$\sum_{i=1}^{n} X_iY_i$

The rest of the computations follow immediately;

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left(n\sum_{i=1}^{n} X^2 - \left(\sum_{i=1}^{n} X\right)^2\right)\left(n\sum_{i=1}^{n} Y^2 - \left(\sum_{i=1}^{n} Y\right)^2\right)}}$$

$$r = \frac{5\times 1013.68 - 102.7\times 49.2}{\sqrt{(5\times 2111.99 - (102.7)^2)(5\times 488.26 - (49.2)^2)}} = \frac{15.56}{\sqrt{12.66\times 20.66}} = 0.96$$

There is a strong positive correlation between x and y

The coefficient of determination

$r^2 = (0.96)^2 = 0.93$

This implies that 93% of the changes in y could be attributed to x if a regression equation of Y on X is desired.

Summary

In this unit, you have learnt how to obtain values for the Correlation and the value must be between 0 and 1.

Self Assessment Questions

(1) Explain correlation

(2) Express how the correlation is obtain from two or more variable.

(3) Show the degree of relationship between the variables whether strong or weak

Tutor Marked Assessment

- Consider the following data, where x denotes the respective number of branches that 10 different banks have in some metropolitan area and y denotes the corresponding share of the total deposits held by the banks:

X	198	186	116	89	120	109	28	58	34	31
Y	22.7	16.6	15.9	12.5	10.2	6.8	6.8	4.0	2.7	2.8

- Consider the following data, where x denotes the average daily temperature in degree Fahrenheit and y denotes the corresponding daily natural gas consumption in cubic feet:

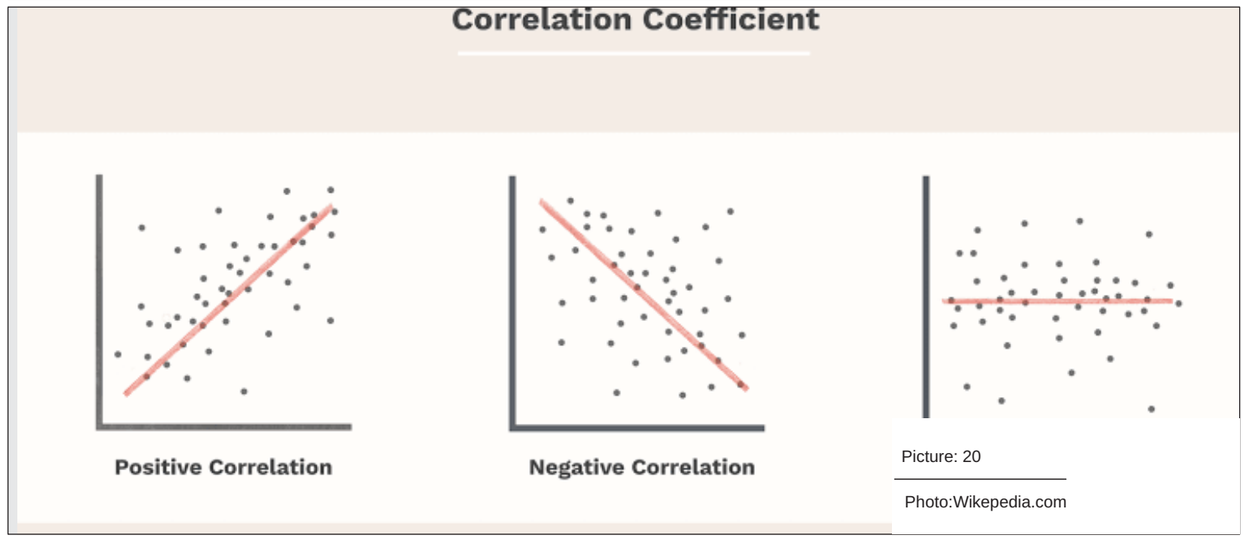
X, ^{0}F	50	45	40	38	32	40	55
Y,ft^3	2.5	5.0	6.2	7.4	8.3	4.7	1.8

- Consider the following data, where x denotes the average daily temperature in degree Fahrenheit and y denotes the corresponding daily stock index average (in 1998)

X	63	72	76	70	71	65	70	74	68	61
Y	8385	8330	8325	8320	8330	8325	8280	8280	8300	8265

References

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and
- Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.
- Gupta (2014). Introduction to Fundamental Statistics.
- Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.
- Introductory Statistics 3rd Edition by Sheldom M. Ross



Picture: 20

Photo:Wikepedia.com

UNIT 3

Continues Correlation

Introduction

We explain what a statistical hypothesis is and show how sample data can be used to test it. We distinguish between the null hypothesis and the alternative hypothesis. We explain the significance of rejecting a null hypothesis and of not rejecting it. We introduce the concept of the p value that results from a test. Tests concerning the mean of a normal population are studied, when the population variance is both known and unknown. One sided and two sided tests are considered.

Learning Outcomes

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

1. Know t – test for correlation coefficient
2. Obtain correlation from two or more variable.
3. Express the degree of relationship between the variables whether strong or weak

Main Content

10 mins

In Testing hypotheses about the population correlation coefficient (r), the joint distribution of X and Y is assumed to be bivariate normal. We will discuss bivariate normal density in this study or work. We can therefore think of a theoretical population coefficient (r). Tests of significance or hypotheses concerning the various values of require the knowledge of the sampling distribution of r. In particular, for = 0 the distribution is symmetrical, and statistic involving the students t- distribution can be used. On the other hand, for = 0 the distribution is skewed, in such case a transformation suggested by Sir, R. A. Fisher produces a statistic that is approximately normally distributed.

The Student's t – Test for Correlation Coefficient

The hypothesis and test statistic for the student's t – test is summarized as follows;

Hypothesis

$H_0 : \rho = 0$

Vs

$H_1 : \rho \neq 0$

Test Statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Or

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Has the students t – distribution with n – 2 degrees of freedom

Decision Criterion

Reject H_0 if $t|t| \geq t_{\frac{\alpha}{2},n-2}$

Fisher's Z – Transformation of Correlation Coefficient

Another versatile test statistic for the correlation coefficient is the Fisher's transformation of the correlation r to a normally distributed random variable denoted by Z_r. The test was introduced by Sir Ronald Aylmer Fisher in 1915. The statistic Z_r is given by

$$Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

In this case, in denotes the natural logarithm of the value in parenthesis.

Note

In any case, we may as well wish to express the Z – statistic in terms of logarithms to base 10, and then we have the following transformation:

$$Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = (1.15131)\log_{10}\left(\frac{1+r}{1-r}\right)$$

For sample from a bivariate normal distribution with sample sizes of 10 or more, the distribution of Z_r is approximately normally distributed with respective mean and variance given by:

$$\xi_0 = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) = \frac{1}{2}\ln\left(\frac{1+0.5}{1-0.5}\right) = 0.5493$$

$$V(Z_r) = \frac{1}{n-3}$$

In this case, n is the sample size

r is the population correlation coefficient

x is the lowercase Greek letter 'Zeta'

The Fisher's r – to - Z_r transformation is remarkably robust. It allows us to use the techniques developed for testing hypotheses about μ (as discussed earlier) to test any hypothesis of the form:

$H_0 : \rho = 0$

Vs

$H_1 : \rho \neq 0$

Or any one –sided alternative and 0r is any specified constant which takes values between -1 and +1 inclusive.

If the null hypothesis is correct, then the expected value of Z becomes;

$$E(Z_0) = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) = \xi_0$$

We know that if a normally distributed random variable X has expected value 0m and variance 2s then we have the standard normal score:

$$Z = \frac{X - \mu_0}{\sigma}$$

Which is distributed N(0,1). Thus by analogy for $\rho = \rho_0$:est statistic becomes,

$$Z = \frac{Z_r - \xi_0}{\sqrt{1/(n-3)}}$$

In most correlational research, the researcher could be more interested in finding a significant correlation. It is therefore understood to mean "significantly different from zero". Hence, the hypothesis reduces to:

$H_0 : \rho = 0$

Vs

$H_1 : \rho \neq 0$

or any one –sided alternative

$$\xi_0 = \frac{1}{2}\ln\left(\frac{1+0}{1-0}\right) = \frac{1}{2}\ln(1) = 0$$

$$Z = \frac{Z_r}{\sqrt{1/(n-3)}} = Z_r\sqrt{n-3}$$

Test Concerning Two Correlation Coefficients

In this unit we have discussed several methods for testing hypotheses about the difference between two population means, 21mm-. The Fisher's r – to - Z_r transformation permits us to somehow extend these methods to testing hypotheses about the difference between two correlation coefficients, 21rr- . Let 1r be the correlation coefficient of X and Y calculated from a sample of 1n observations on population 1 and let 2rbe that he correlation coefficient of X and Y calculated from a sample of 2n observation on population 2. Then from the properties mathematical expectations, we have

$$E(Z_{r_1} - Z_{r_2}) = \xi_1 - \xi_2 = \frac{1}{2}\ln\left(\frac{1+\rho_1}{1-\rho_1}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_2}{1-\rho_2}\right)$$

Based on the assumption that samples drawn from distinct populations are independent, we have

$$V(Z_{r_1} - Z_{r_2}) = \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}$$

Therefore, if the joint distribution of X and Y is bivariate normal and if 1*n* and 2*n* are both greater than 10, then the statistic Z is approximately normally distributed N(0,1) as follows;

$$Z_{r_1 - r_2} = \frac{(Z_{r_1} - Z_{r_2}) - (\xi_1 - \xi_2)}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

If the null hypothesis

$H_0 : \rho_1 = \rho_2$: then $\xi_1 - \xi_2 = 0$ and the test statistic reduces to

$$Z_{r_1 - r_2} = \frac{(Z_{r_1} - Z_{r_2})}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{\frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right) - \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right)}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

In practice, this is the most widely applied test for testing the significant difference between two correlation coefficients. Hence, we shall limit our discussion to testing the significant difference between two correlation coefficients as follows:

$H_0 : \rho = 0$

Vs

$H_1 : \rho \neq 0$

or any appropriate one – sided alternative and the test statistic becomes;

$$Z = \frac{(Z_{r_1} - Z_{r_2})}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

or

$$Z = \frac{\frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right) - \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right)}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Spearman's Rank Correlation Coefficient

Since the underlying assumptions for the significance test for the product – moment correlation coefficient are rather stringent, it is sometimes preferable to use a nonparametric alternative. Most popular among such nonparametric measures of association is the rank correlation coefficient, also called the Spearman's rank correlation coefficient. The Spearman's rank correlation was first introduced by an English statistician, Charles Edward Spearman in 1904. The rank correlation is a common measure of association for random variables X and Y; where the X and Y observations are replaced by their respective ranks. Hence, the rank correlation was derived from the product – moment correlation by replacing the actual observations by their ranks. The Spearman's rank correlation coefficient is given by the formula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

In this case, $d_i = r_{x_i} - r_{y_i}$ is the respective difference in ranks between X and Y

r_{x_i} is the ranks of x observations

r_{y_i} is the ranks of y observations

n number of observations

and $-1 \leq r_s \leq +1$

Derivation of the Rank Correlation

Consider the product – moment correlation below;

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)\left(n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right)}}$$

Steps Involved in Calculating the Rank Correlation

(1) Rank the respective paired observations in ascending (or descending) order

(2) Pair the corresponding ranks of the two variable

(3) Obtain the difference of the paired ranks d_i

(4) Square the difference of the ranks and sum them up

(5) Substitute the sum of the square difference and n into the formula of the rank correlation and evaluate.

Illustration

Example 1

The following table shows the heights of husbands and the corresponding height of their wives, both in cm, in a random sample of ten small families.

Husband's height	Wife's height
165	173
160	168
170	173
163	165
173	175
158	168
178	173
168	165
173	180
170	170
165	173
160	168

(a) Calculate the correlation coefficient between x and y

(b) Conduct a test of significance for the correlation coefficient at the 0.05 level of significance

Solution

(a) The product – moment correlation coefficient

n = 12

$$\sum_{i=1}^{n} X_i = 1678$$

$$\sum_{i=1}^{n} Y_i = 1710$$

$$\sum_{i=1}^{n} X_i^2 = 281924$$

$$\sum_{i=1}^{n} Y_i^2 = 292610$$

$$\sum_{i=1}^{n} X_i Y_i = 287103$$

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)\left(n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right)}}$$

$$r = \frac{12\times 287103 - 1678\times 1710}{\sqrt{\left(12\times 281924 - (1678)^2\right)\left(12\times 292610 - (1710)^2\right)}} = \frac{1620}{\sqrt{3556\times 2000}} = 0.620$$

(a)The test of significance

Hypothesis

$H_0 : \rho = 0$ (The correlation coefficient is not significant)

Vs

$H_1 : \rho \neq 0$ (The correlation coefficient is significant)

Level of significance

$\alpha = 0.05$

Test Statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Decision Criterion

Reject H_0 if $|t| \geq t_{0.025,8} = 2.306$

Computations

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.62\sqrt{\frac{12-2}{1-(0.62)}} = 2.24$$

$|t| = 2.24$

Conclusion

Since t=24.2 does not exceed $t_{0.025,10} = 2.306$ then H_0 cannot be rejected. Hence, the correlation is not statistically significant.

Example 2

Ten adults underwent a physical training programme. After the programme, their weights (in Kg) and anaerobic threshold (in liters/min) were measured, with the results below:

Husband's height	Wife's height
62	1.37
57	1.34
94	1.93
69	1.92
66	2.24
76	2.02
58	1.35
88	2.21
70	1.79
84	1.74
62	1.37
57	1.34

You are hereby required to;

(a) Compute the product – moment correlation coefficient

(b) Use the Z – transformation to test the hypothesis, at the 1% level of significance, that the population correlation coefficient is greater than 0.5.

Solution

(a) The product – moment correlation coefficient

n = 12

$$\sum_{i=1}^{n} X_i = 724$$

$$\sum_{i=1}^{n} Y_i = 17.91$$

$$\sum_{i=1}^{n} X_i^2 = 53886$$

$$\sum_{i=1}^{n} Y_i^2 = 33.1201$$

$$\sum_{i=1}^{n} X_i Y_i = 1320.82$$

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)\left(n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right)}}$$

$$r = \frac{12\times 1320.82 - 724\times 17.91}{\sqrt{(12\times 53886 - (724)^2)(12\times 33.1201 - (17.91)^2)}} = \frac{241.36}{\sqrt{14684\times 10.4329}} = 0.620$$

(a) The test of the correlation coefficient using the Z – transformation

Hypothesis

$H_0 : \rho = 0$ (The correlation coefficient is 0.5)

(Vs)

$H_1 : \rho \neq 0$ (The correlation coefficient exceeds 0.5)

Level of significance

α = 0.01

Test Statistic

$$Z = \frac{Z_r - \xi_0}{\sqrt{1/(n-3)}}$$

Decision Criterion

Reject H_0 if t $Z \geq Z_{0.01} = 2.33$

Computation

$$Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0.62}{1-0.62}\right) = 0.7250$$

$$Z = \frac{Z_r - \xi_0}{\sqrt{1/(n-3)}} = \frac{0.7250 - 0.5493}{\sqrt{1/7}} = \frac{0.1757}{0.3780} = 0.465$$

Z = 0.47

Conclusion

Since Z = 0.47 does not exceed $Z_{0.01} = 2.33$ then H_0 cannot be rejected.

Hence, the correlation coefficient does not exceed 0.5

Example 3

A thoracic cardiology claims that there is higher correlation between age (in years) and blood pressure (in millimeter of mercury) among male adults than female adults. Suppose that r_1= 0.75 is correlation coefficient of age (X) and blood pressure (Y) of 18 male adults. Suppose further that r_2 = 0.69 is correlation coefficient of age (X) and blood pressure (Y) of 20 female adults. You are required to test, at the 5% level of significance, either male adults actually have higher correlation between their age (X) and blood pressure (Y) as claimed by the cardiologist.

Solution

Hypothesis

$H_0 : \rho = 0$ (There is no significant difference between male and female)

Vs

$H_1 : \rho \neq 0$ (The male correlation exceeds the female correlation)

Level of significance

$\alpha = 0.05$

Test Statistic

$$Z = \frac{(Z_{r_1} - Z_{r_2})}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Decision Criterion

Reject H_0 if t $Z \geq Z_{0.05} = 1.645$

Computation

$r_1 = 0.75$

$r_2 = 0.69$

$n_1 = 18$

$n_2 = 20$

$$Z_{r_1} = \frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right) = \frac{1}{2}\ln\left(\frac{1+0.75}{1-0.75}\right) = 0.9730$$

$$Z_{r_2} = \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right) = \frac{1}{2}\ln\left(\frac{1+0.69}{1-0.69}\right) = 0.988480$$

$$Z = \frac{(Z_{r_1} - Z_{r_2})}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0.9730 - 08480}{\sqrt{\frac{1}{15} + \frac{1}{17}}} = \frac{0.1250}{0.3542} = 0.353$$

Z = 0.353

Conclusion

Since Z = 0.353 does not exceed $Z_{0.05}$ = 1.645 then H_0 cannot be rejected. Hence, there is no significant difference male and female correlation between their age and blood pressure. Therefore, there is a reason to doubt the cardiologist's claim.

Summary

In this unit you have learnt:

- How to obtain values for the Correlation coefficient and the value must be between 0 and 1.

Self Assessment Questions

(1) Know t – test for correlation coefficient

(2) Obtain correlation from two or more variable.

(3) Express the degree of relationship between the variables whether strong or weak

Tutor Marked Assessment

- The following tables shows the number of people infected with malaria in thousand (x) and the number of clinics per 100,000 population (y) of a random sample of 12 major cities in Nigeria. You are required to calculate the coefficient of rank correlation between x and y.

X	110	100	100	94	105	97	94	93	99	93	90	86
Y	12	18	24	37	52	64	69	81	98	101	106	119

- The following table shows the scores of eight pairs of real twins on intelligence tests. The goal is to see if there is underlying linear relationship between the test scores of the first – contained inborn twin and that of the second – born twin. The highest scores corresponding to the best result as contained in the data shown in the table below.

First – born (x)	90	75	99	60	72	83	83	90
Second - born (y)	88	79	98	66	64	83	88	98

- The following tables shows the number of vehicles in millions (x) and the number of road accidents in thousand (y) of a random sample of 10 big cities in Nigeria.

X(in Millions)	Y (in 1000)
2.6	138
3.1	163
3.5	166
3.7	153
4.1	177
4.4	201
4.6	216
4.9	208
5.3	226
5.8	238

Use the table to:

- Calculate the coefficient of rank correlation between x and y
- Conduct a test of significance for the rank correlation at the 5% level of significance.

References

Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and

Aliyu Usman (2012). Statistical Methods for Biometric and Medical Research.

Gupta (2014). Introduction to Fundamental Statistics.

Introductory Statistics 3rd Edition by Openstex College Adapted by Mark Beintema and Natalia Casper.

Introductory Statistics 3rd Edition by Sheldom M. Ross