Statistics – iGAP Technologies Pvt. Ltd

Introduction

• What is Statistics?

• Statistics is the science of conducting studies to collect, organize, summarize, analyse, and draw conclusions from data.

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and, organization of data. The word Statistics is derived from the Latin word Status, which is loosely defined as a statesman.

Example of Statistics:

- It was reported that violent crimes were down by 3.5% in 2010 in the world. It was reported that the average student loan debt was about $28,000.
- The College stress and Mental illness poll reported that 85% of college and university students reported feeling stress daily; 75% reported stress from school work, and 64% experienced stress from grades.

Why we should study Statistics?

Statistics is the science and also the art of learning from data. As a discipline it is concerned with the collection, analysis, and interpretation of data, as well as the effective communication and presentation of results relying on data. Statistics lies at the heart of the kind of quantitative reasoning necessary for making important advances in the sciences, such as medicine and genetics, and for making important decisions in business and public policy.

Variable, Data, Population, Sample

• 'A variable is a characteristic or attribute that can assume different values.

• The values that a variable can assume are called data.

• A population consists of all subjects (human or otherwise) that are studied, (all members of a defined group that we are studying or collecting information on for data driven decisions).

• A sample is a subset of the population.(A part of the population is called a sample).

Random Variables.

• Variables whose values are determined by chance are called random Variables.

• Example. Automobile insurance, claim suppose 5% every year.

• Population: Census example • Sample: may be biased.

Descriptive and Inferential Statistics

Statistics is divided into two main areas, depending on how data are used.

• Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

• Inferential statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

Descriptive statistics

• In descriptive statistics the statistician tries to describe a situation.

• Descriptive statistics give information that describes the data in some manner. For example, suppose a grocery store sells Eggs, Bread, Milk and fruit. If 100 items are sold and 30 out of the 100 were Milk, then one description of the data on the grocery store items sold would be that 30% were Milk.

• This same grocery store may conduct a study on the number of bread sold each day after one month and determine that an average of 20 Bread were sold each day. The average is an example of descriptive statistics.

• Another example, consider the national census conducted by any country in every 10 years. Results of this census give you the average age, income, gender and other features of the population. To obtain this information, the govt must have some mean to collect significant data. Once the data are collected, they organize and summarize Full sere them. Finally, they present the data in some meaningful form, such as chart, reports graph and table etc.

Descriptive statistics cont.

• A graphical representation of data is another method of descriptive statistics. Examples of this visual representation are histograms, bar

graphs and pie graphs etc. Using these methods, the data is described by compiling it into a graph, table or other visual representation.

Inferential Statistics

• Inferential statistics makes inferences about populations using data draw from the population. Instead of using the entire population to gather the data, the statistician will collect samples from the millions of residents and make inferences about the entire population using the sample. The sample is a set of data taken from the population to represent the population. Probability distributions, hypothesis testing, correlation testing and regression analysis all fall under the category of inferential statistics.

• In inferential statistics, the answers are never 100% accurate because calculations use a sample taken from the population. This sample does not include every measurement from the population.

Lecture 2 – variable and type of data

Recorded Values and Boundaries

Variable	Recorded Value	Boundaries
Length	18 centimetres	17.5-18.5 cm (cm)
Temperature	76° Fahrenheit	75.5-76.5 °F (°F)
Time	0.24 second (sec)	0.235-0.245 sec
Mass	3.8 gram	3.75-3.85 gram

• How variables are categorized counted or measured.

• Example. Can the data values be ranked, 1s place, 2nd place and so on

• Can the data be organized into specific categories (rural, urban, etc.) • Can the data be measured (height, temperature, time, length, IQ's etc.) • these types of classification needs measurement scale.

4 common types of scales

These are simply ways to categorize different types of variables.

Nominal - categorical (names), No ranking or no order.

Statistics – iGAP Technologies Pvt. Ltd

Nominal scales are used for labelling variables, without any quantitative value. Examples, Note that all of these scales are mutually exclusive (no overlap) and none of them have any numerical significance. No ranking or no order.

What is your gender?

• M - Male

F - Female*

What is your hair colour?

• 1 – Brown

2 - Black

3 – Blonde

4 - Gary Other

• Data measured at this level can be placed into categories, and these categories can be ordered or ranked.

• Example: Letter Grades A, B, C, and D...,

• Guest speaker speech, excellent, average, poor

• Restaurant services. "1" for poor, "2" for average, "3" for very good and "4" for excellent. Students ranking 1st, 2nd 3rd,

IQ score: meaningful difference between 107 and 108 IQ score. Temperature: meaningful difference between 15C and 16C

• Distance: meaningful difference between 30km-40km. precise Note: there is no true ZERO. IQ tests do not measure people who have no intelligence.

• Temperature OC does not mean no heat at all. Etc.

Ratio - interval, plus ratios are consistent, true zero

• The zero in the scale makes this type of measurement unlike the other types of measurement, although the properties are similar to that of the interval level of measurement.

• Example: it is used to measure height, phone calls received, area, weight etc. if one person weight is 100 pounds. While other is 50 pounds, then the ratio is 2:1

• The researcher should note that among these It measurement, the nominal level is simply used data, whereas the levels of measurement described by 25s interval level and the ratio level are much more exact.

3 Data Collection and Sampling Techniques

Why Statisticians Collect data?

• Data can be used to describe situations or events.e.g

• Manufacturer can make a smart marketing strategy if he knows the purchasing power of the consumers. With the help of data buyers can make an intelligent decisions, what stock to buy etc.

For these purposes statisticians need data. There are different ways to collect data. The most common way is to collect data through surveys.

• Telephone Survey

Mailed questionnaire &

Personal Interviews

Telephone Surveys

Advantages

• Telephone surveys are less expensive.

• People are more frank, to express their judgement, as no face to face communication.

Disadvantages

• Not all people have chanced to survey, as some times they don't receive the phone, or they are at work, when the call was made. Some people have unlisted numbers, or don't have phone.

Mailed Questionnaire

Advantages

• Mail survey are less costly than personal survey, and can cover the wide area than telephone survey.

People may remain anonymous if they want.

Disadvantages

• Low number of replies.

Incorrect responses.

• May trouble to read or understand questions. Etc.

Personal Survey/Interview

Advantages

Detail answer from the respondents.

Disadvantages

• More Costly.

• Unfair choice of respondents.

• Researchers use sample to collect data, which saves time & money, but sample may be unfair or biased.

• To get sample that are unbiased, that give each subject in the population an alike chance of being selected, statisticians practice four methods of sampling.

Four methods of Sampling

• **Random** - random number generator

• Random sample is a subset of the population in which each member of the subset has an equal probability of being selected. An example of a random sample would be the names of 30 employees being chosen out of a hat from a company of 300 employees.

• **Systematic** - every kth subject

• Sample members from a larger population are selected according to a random starting point and a fixed periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size.

• **Stratified** - divide population into "layers"

• Stratified random sampling is a method of sampling that involves the division of a population into smaller groups known as strata

• For example, one might divide a sample of adults into subgroups by age, like 17-28, 29- 38, 39-48, 49-58, etc.

Cluster - use intact groups

• With cluster sampling, the researcher divides the population into separate groups, by some means such as geographic area or school zone etc. called clusters. Then randomly choose some of these clusters as a whole as a subjects of the samples.

Other sampling methods

• Convenient - mall surveys

• Convenience sampling (also known as availability sampling) is a particular kind of non-probability sampling technique that depend on data collection from population members who are conveniently available to contribute in study. Facebook polls or questions can be stated as a common example for convenience sampling.

• Sampling error - Sample vs. population

• A sampling error is a statistical error that arises when an analyst does not select a sample that characterises the whole population of data and the results found in the sample do not signify the results that would be acquired from the entire population. • Or • the difference between the results found from a sample and the results found from the population from which the sample was selected. • Non sampling error • A non-sampling error is a statistical error triggered by human blunder to which an exact statistical analysis is exposed. These errors can include, but are not limited to, data entry errors, biased questions in a questionnaire, unfair processing/decision making, unsuitable analysis and Wrong information.

Lecture 4 -

4 difference between observational and experimental study

Observational Studies

• An observational study is a study where researchers simply collect data based on what is seen and heard and conclude based on the data collected. Or

• Researchers merely observe what is happening or what has happened in the past and tries to draw inferences based on these observations.

Experimental Studies

An experimental study is a method of applying treatments to a group and recording the effects. In an experiment, you control the conditions. Or

Statistics – iGAP Technologies Pvt. Ltd

Researchers manipulates one of the variables and tries to determine how the manipulation influences other variables.

Example – traffic signal.

Lecture 5 - concept of probability

Probability –

The probability is the number of favourable outcomes, out of the number of possible outcomes.

• Flipping a coin has two possibilities: heads or tails.

• P(Heads) = 1/2 or 50%

• P (Tails) = 1/2 or 50%

Examples OF Probability:

• You roll a 6-sided die.

• What is P(greater than 5)?

• (greater than 5) = ?

• The die has 6 sides, numbered 1, 2, 3, 4, 5 and 6 .

There is 1 number greater than 5

=1/6 Ans.

• You pick a number at random

- What is P(prime)?

Write your answer as a percentage

P(prime) =?

• There are 4 numbers. Numbered 7,8,5 and 4. the prime number are 7 and 5

• =2/4

= .5 or 50%

• You roll a 6-sided die. What is P(prime)?

Write your answer as a percentage. P (prime) =?

The only factors of a prime number are 1 and itself. The number 1 is neither prime nor composite. The die has 6 sides numbered 1,2,3, 4, 5 and 6. The prime no are 2,3 and 5.

P(prime) = 3/6

=0.5

=50%

Same as even no example.

Lecture 6 - raw data grouped frequency distribution categorical frequency distribution

Raw Data and Organizing Data

Data collected in original form is called raw data. (Primary data or source data)

Example – collection of pages maths, physics, chemistry.

Mine.

• Raw data refers to any data object that hasn't undergone through processing, either manually or through computer software.

• A frequency distribution is the organization of raw data in table form, using classes and frequencies. Or

• An arrangement of statistical data that exhibits the frequency of the occurrence of the values of a variable.

Ages of the 54 wealthiest people in the world. (Just assume, the data is not real) ¢ 54 34 45 56 67 81 45 23 34 56 76 54 32 82 84 34 56 45 76 34 24 23 85 27 76 53 44 37 82 56 78 57 41 83 48 46 58 59 63 64 76 58 42 41 35 65 45 78 70 60 50 40 35 50

Class Limit		Tally	Frequency
23 - 29	w-7	||||	4
30-36		||||| ||	7
37 - 43			
44 - 50			
51 - 57			
58 - 64			

65 - 71

72 - 78

79-85 ||||| | 6;

Basic rules for classes:

• There should be between 5 and 20 classes.(no hard and fast rule)

• The classes must be mutually exclusive. This means that no data value can fall into two different classes.

• The classes must be all inclusive or exhaustive. This means that all data values must be included.

• The classes must be continuous. There are no gaps in a frequency distribution. Classes that have no values in them must be included (unless it's the first or last class which are dropped).

• The classes must be equal in width. The exception here is the first or last class. It is possible to have an "below " or " and above class. This is often used with ages.

2 types of frequency distributions

•There are 2 types of frequency distributions that are most often used.

• Categorical frequency distributions

• Grouped frequency distributions

Categorical frequency distributions

• Nominal- or ordinal-level data that can be placed in categories is organized in categorical frequency distributions. Or

• A frequency distribution in which the data is only nominal or ordinal is called categorical frequency distributions.

• M&M colours, Religion, Political parties, Blood type etc.

Categorical frequency distribution

Suppose 30 patients blood was taken to determine their blood type, and the data is as follows.

A, B,B,AB,O 0,A,B,AB,B B,B,0,A,0 A,0,0,0,AB AB,A,O,B,A A,O,B,AB,AB Raw data, we have to arrange this data in a meaningful manner.

A Class	B Tally	C Frequency	D Percent (%=£/n*100)
A		7	=7/30*100 23.33%
B		8	=8/30*100 26.66% 30%
AB		9	30%
O		6	20%
		Total 30	

We conclude more people have blood o group type.

Grouped Frequency Distributions

When the range of the data is large, then we use grouped frequency distribution. Grouped frequency distribution is the organizing of raw data in table form, using classes and frequencies. The largest data value that can be included in a class is the upper class limit for that class; the smallest data value that can be included is the lower class limit.

• Class boundaries separated the class.

Creating a Grouped Frequency Distribution

• Find the largest and smallest values.

• Compute the Range = Maximum - Minimum. Select the number of classes desired. This is usually between 5 and 20. Find the class width by dividing the range by the number of classes and rounding up

• To find the upper limit of the first class, subtract one from the lower limit of the second class. Then continue to add the class width to this upper limit to find the rest of the upper limits. * Find the boundaries by subtracting 0.5 units from the lower limits and adding 0.5 units from the upper limits. The boundaries are also half-way between the upper limit of one class and the lower limit of the next class. Depending on what you're

trying to accomplish, it may not be necessary to find the boundaries.
•Tally the data.

•Find the frequencies.

• Find the cumulative frequencies. Depending on what you're trying to accomplish it may not be necessary to find the cumulative frequencies. • If necessary, find the relative frequencies and /or relative cumulative frequencies.

Record ages of 54 people.(Just assume, the data is not real, we'll take 9 classes.) 54 34 45 56 67 81 45 23 34 56 76 54 32 82 84 34 56 45 76 34 24 23 85 27 76 53 44 37 82 56 78 57 41 83 48 46 58 59 63 61 76 58 42 41 35 65 45 78 70 60 50 40 35 50

Range =85-23=62

width =62/9 4.88

approx. 7 width

Class Limit

23 - 29

30 - 36

37-43

44 - 50

51-57

58 - 64

65 - 71

72 - 78

79 - 85

• We will choose the lowest data value, 23, for the first lower class limit.

• The subsequent lower class limits are found by adding the width to the previous lower class limits.

•The first upper class limit is one less than the next lower class limit.

• The subsequent upper class limits are found by adding the width to the previous upper class limits .

Ungrouped frequency distributions.

Ages of 28 children, from 10 to 15 years.

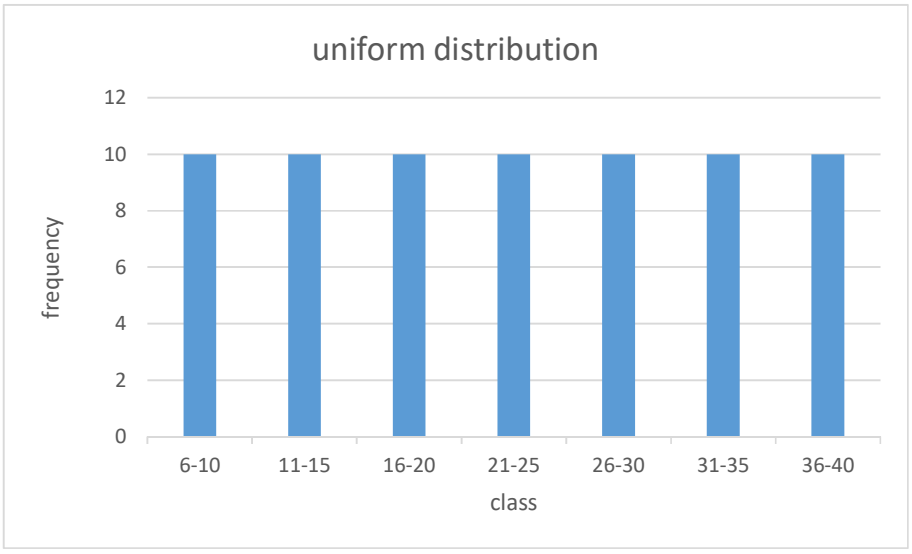
Class Limit	Class Boundaries	freq
10 years	9.95-10.5	4
11	10.5-11.5	
12	11.5-12.5	
13	12.5-13.5	
14	13.5-14.5	
15	14.5-15.5	5

Lecture 7

Uniform distribution – notebook

A uniform distribution, also known as flat or rectangular distribution, is a distribution that has constant probability.

class	frequency
6-10	10
11-15	10
16-20	10
21-25	10
26-30	10
31-35	10
36-40	10



Statistics – iGAP Technologies Pvt. Ltd

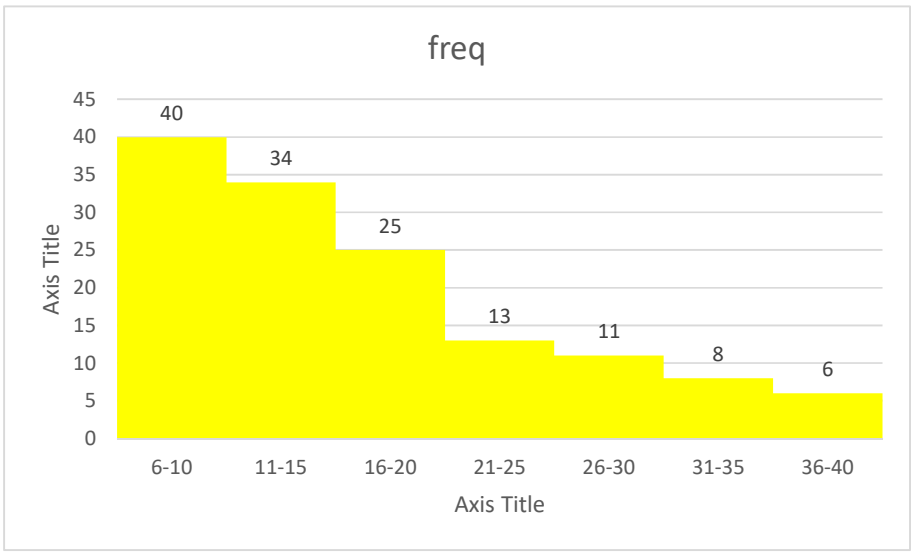
Lecture 8

Positively right skewed – notebook

A "skewed right" distribution is one in which the tail is on the right side. The peak of the distribution is to the left, and data values decreases towards right.

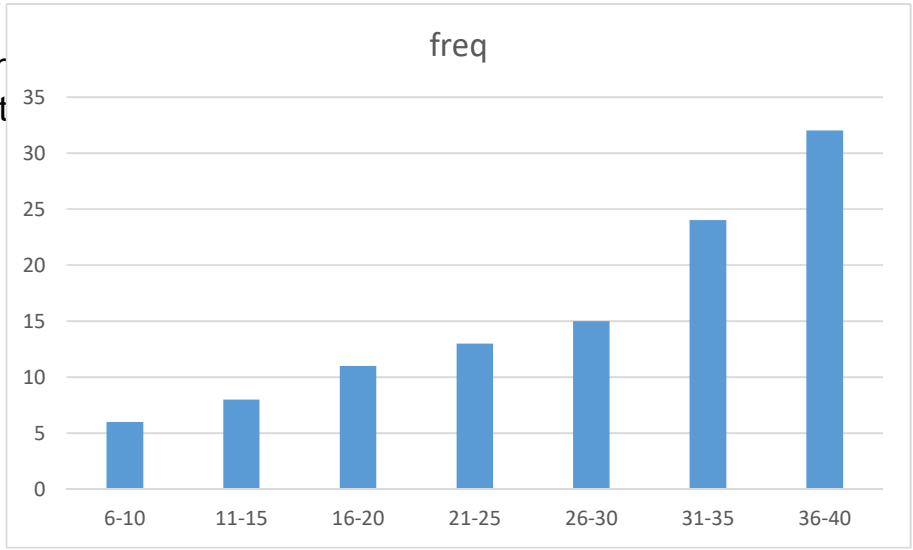
positively or right skewed

class	freq
6-10	40
11-15	34
16-20	25
21-25	13
26-30	11
31-35	8
36-40	6



Negatively right skewed –

A "skewed left" distributior
peak of the distribution is t
towards left.



class	freq
6-10	6
11-15	8
16-20	11
21-25	13
26-30	15

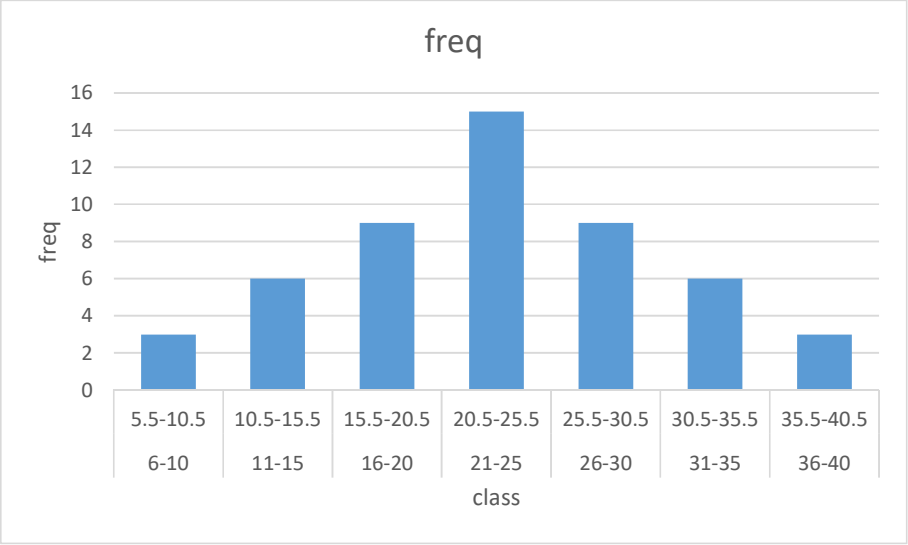
31-35	24
36-40	32

Lecture 9

Bell shaped distribution –

The term bell curve is used to describe the mathematical concept called normal distribution. A bell curve is another name for a normal distribution curve or Gaussian distribution. A bell curve is the most common type of distribution for a variable, and due to this fact, it is known as a normal distribution. Bell shape distribution has single peak and tapers off at either end. It is approximately symmetric i.e. It is roughly the same on both sides.

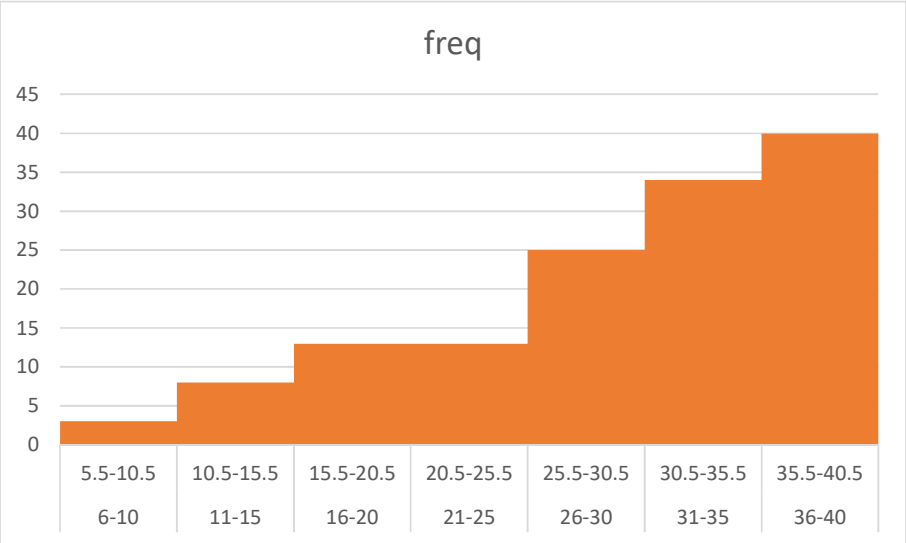
class	class boundries	freq
6-10	5.5-10.5	3
11-15	10.5-15.5	6
16-20	15.5-20.5	9
21-25	20.5-25.5	15
26-30	25.5-30.5	9
31-35	30.5-35.5	6
36-40	35.5-40.5	3



Statistics – iGAP Technologies Pvt. Ltd

Lecture 10 – J shaped distribution

It has more data values on right side, and decreases towards left side.



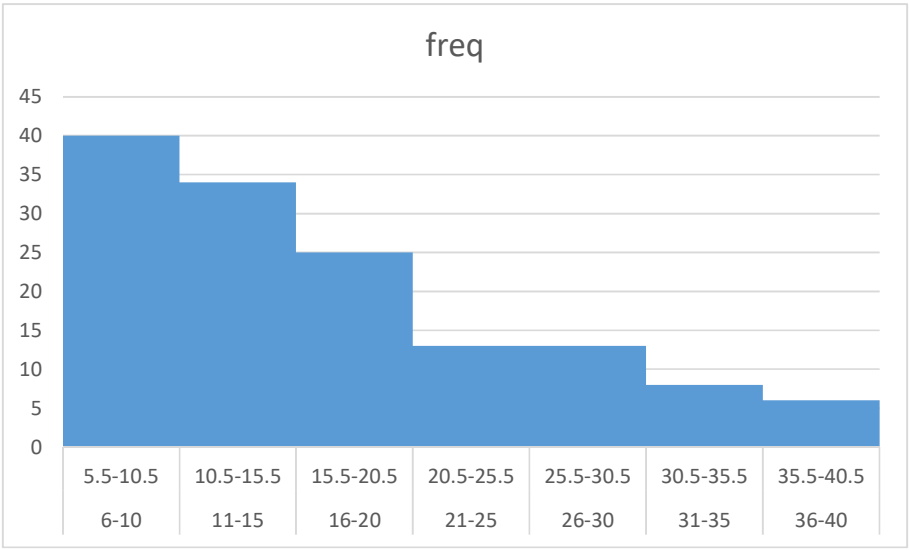
class	class boundries	freq
6-10	5.5-10.5	3
11-15	10.5-15.5	8
16-20	15.5-20.5	13
21-25	20.5-25.5	13
26-30	25.5-30.5	25
31-35	30.5-35.5	34
36-40	35.5-40.5	40

Reverse j shaped dist

It has more data values on left side, and decreases towards right side.

class	class boundries	freq
6-10	5.5-10.5	40
11-15	10.5-15.5	34
16-20	15.5-20.5	25

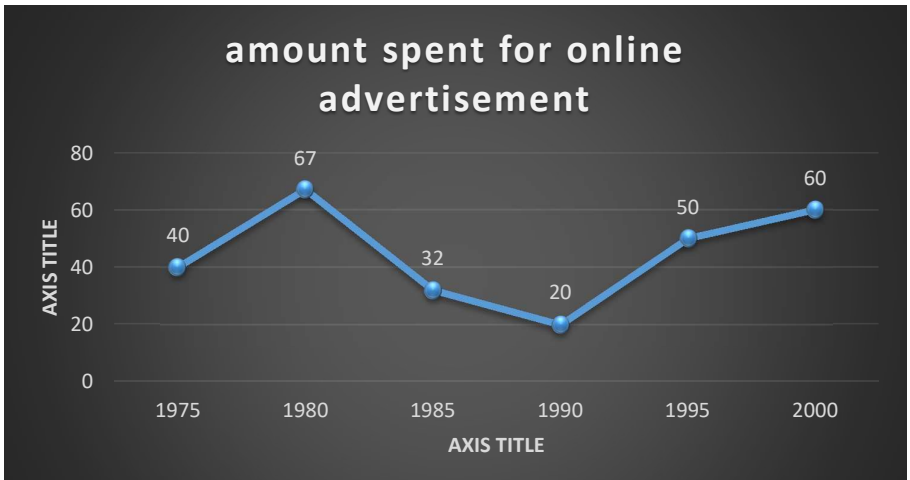
21-25	20.5-25.5	13
26-30	25.5-30.5	13
31-35	30.5-35.5	8
36-40	35.5-40.5	6



Lecture 11 – time series chart/graph

A time series chart, also called a times series graph or time series plot, is a data visualization tool that explains data points at successive intervals of time.

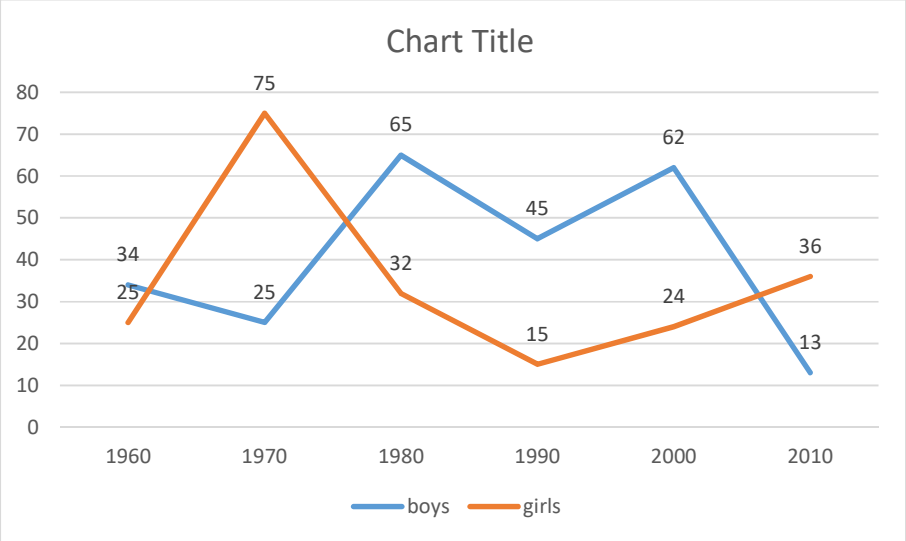
Year	amount spent for online advertisement
1975	40
1980	67
1985	32
1990	20
1995	50
2000	60



Compound time series chart –

Two or more data sets can be compared on the same graph.

Year	boys	girls
1960	34	25
1970	25	75
1980	65	32
1990	45	15
2000	62	24
2010	13	36



Lecture 12 – pair data and scatter plot

A scatter plot (also called a scatter graph, scatter chart, or scatter diagram). The Scatter Diagram graphs pairs of numerical data to look for a relationship between them. They have a very specific purpose. Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation.

Independent variable on x axis and dependent variable on y axis.

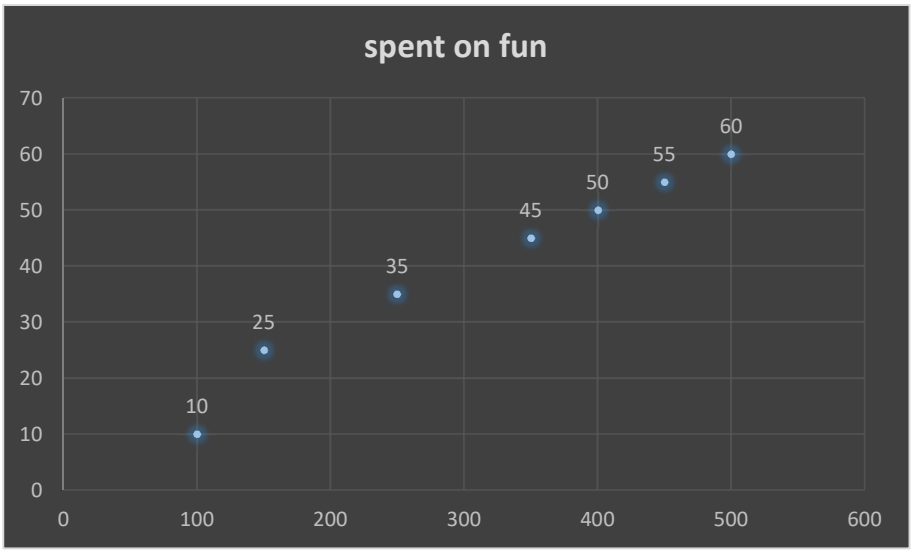
Positive linear relationship

Statistics – iGAP Technologies Pvt. Ltd

A linear relationship means to represent the relationship between two sets of variables with a line (the word "linear" means "a line*.

In other words, a linear line on graph is where you can see a straight line with no curves.

Value y increases as x increases.



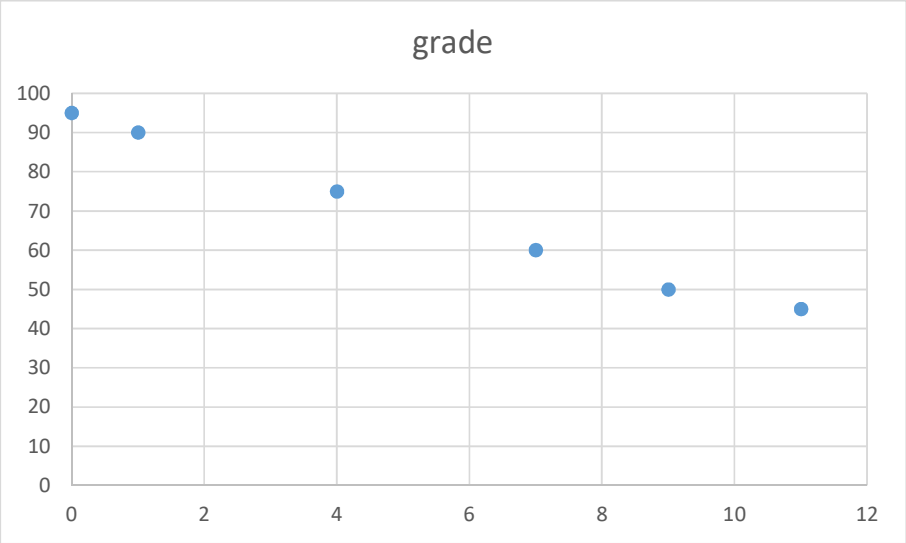
Negative linear relationship –

Negative correlation is a relationship between two variables in which one variable increases as the other decreases, and vice versa a perfect negative correlation is represented by value -1.00 while a 0.00 indicates no correlation and a +1.00 indicates perfect positive correlation.

Independent variable on x axis and dependent variable on y axis.

absence of class	grade

9	50
11	45
1	90
0	95
7	60
4	75



Nonlinear relationship –

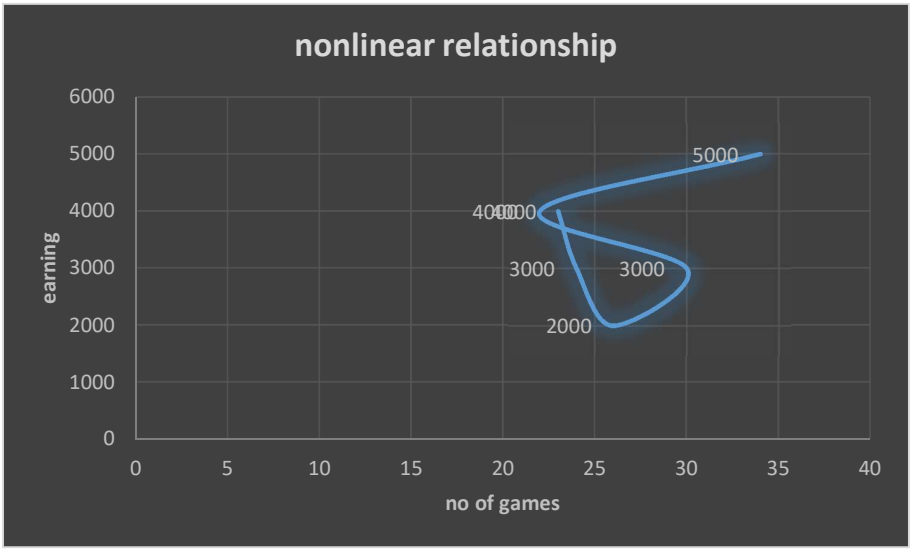
A nonlinear relationship is type of relationship between two entities in which change in one entity does not correspond with constant change in other entity.

The graph of this relationship will be curve instead of straight line.

Independent variable on x axis and dependent variable on y axis.

Number of games	Earning
34	5000
22	4000
30	3000
26	2000
24	3000
23	4000

Statistics – iGAP Technologies Pvt. Ltd

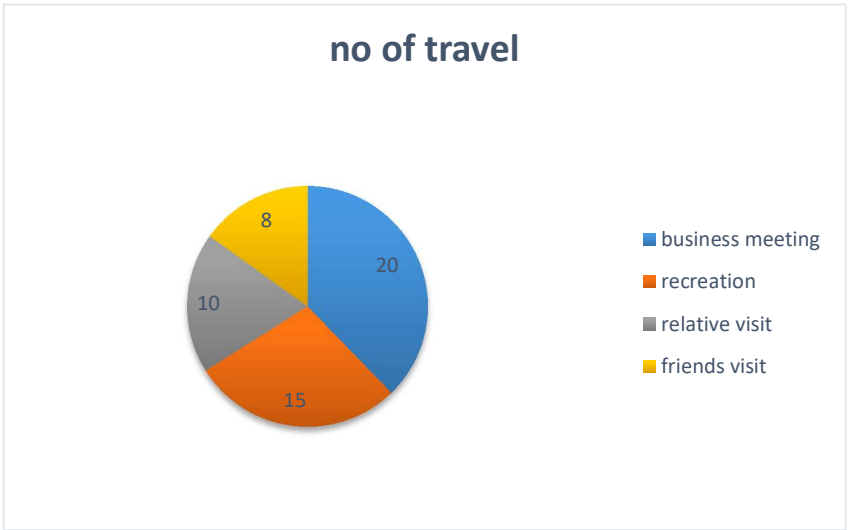


Lecture 13 – pie chart

A circle graph/pie graph is a means of summarizing a set of categorical data or displaying the different values of a given variable (e.g. percentage distribution).

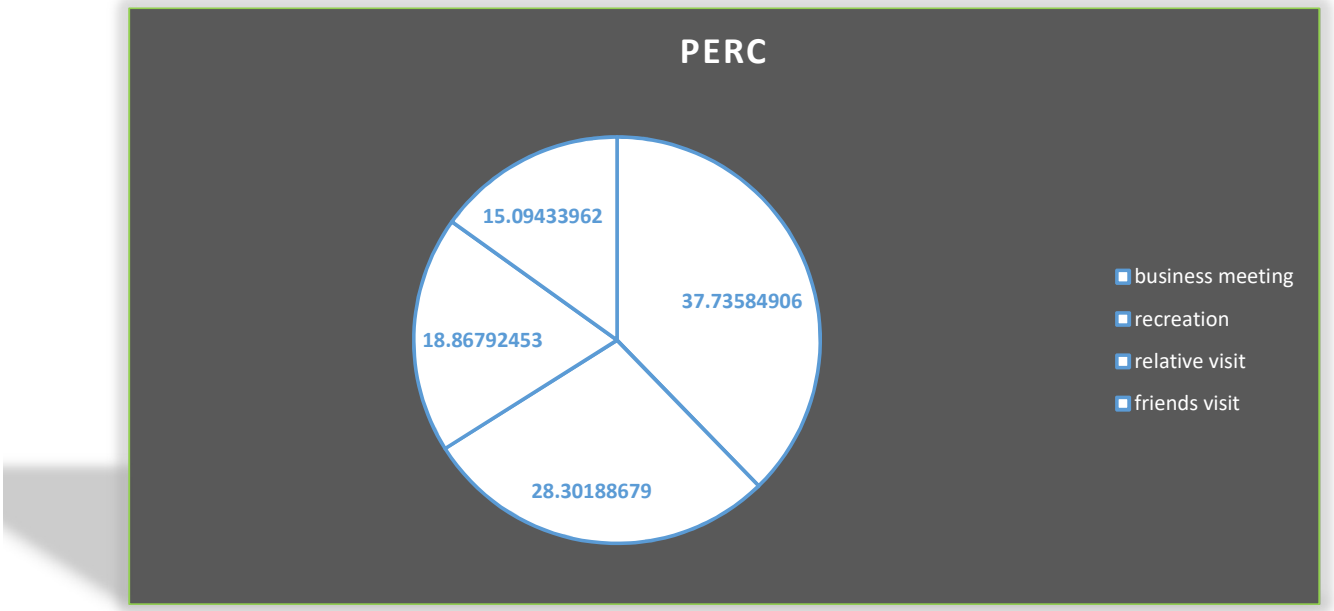
This type of chart is a circle divided into a series of segments/slices. Each slice represents a particular category.

reasons	no of travel
business meeting	20
recreation	15
relative visit	10
friends visit	8
total	53



Statistics – iGAP Technologies Pvt. Ltd

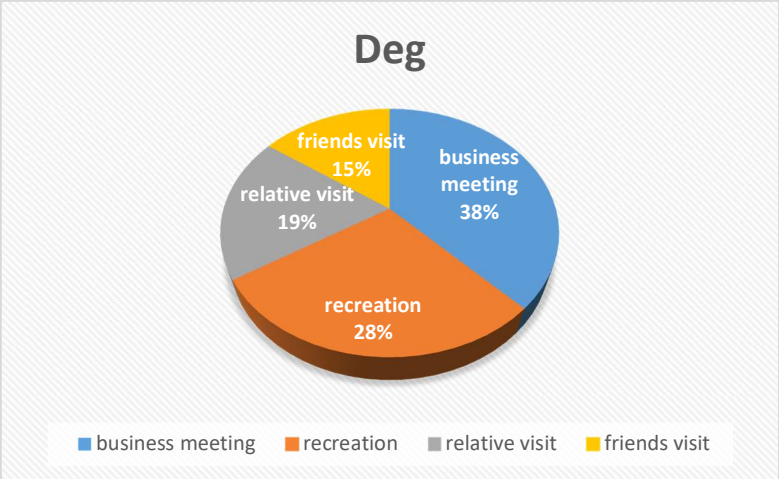
reasons	no of travel	percentage
business meeting	20	37.73585
recreation	15	28.30189
relative visit	10	18.86792
friends visit	8	15.09434
total	53	100



There are 360° in a circle, so the frequency for each class must be converted to a proportional part of a circle.

Degree = f/n °360°

reasons	no of travel	Degree
business meeting	20	135.8491
recreation	15	101.8868
relative visit	10	67.92453
friends visit	8	54.33962
total	53	360



Lecture 14 – Dotplots

• Dot plots are a visualization of data which can give information about the frequency distribution of the data.

A dot plot is a statistical graph/ chart in which each data value is plotted as a dot above the horizontal axis.

If the data value occurs more than once, the corresponding points are plotted above one another.

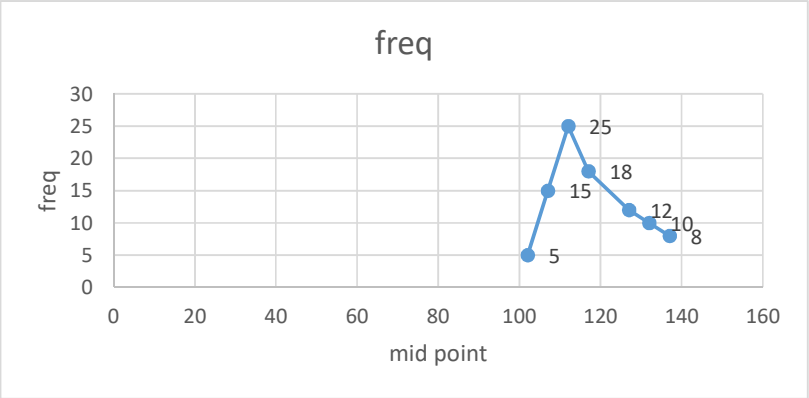
Example – notebook.

Lecture15 – frequency polygon and relative frequency polygon.

Frequency polygons use class midpoints and frequencies of the classes. Lines for the frequency polygon begin and end on the x axis.

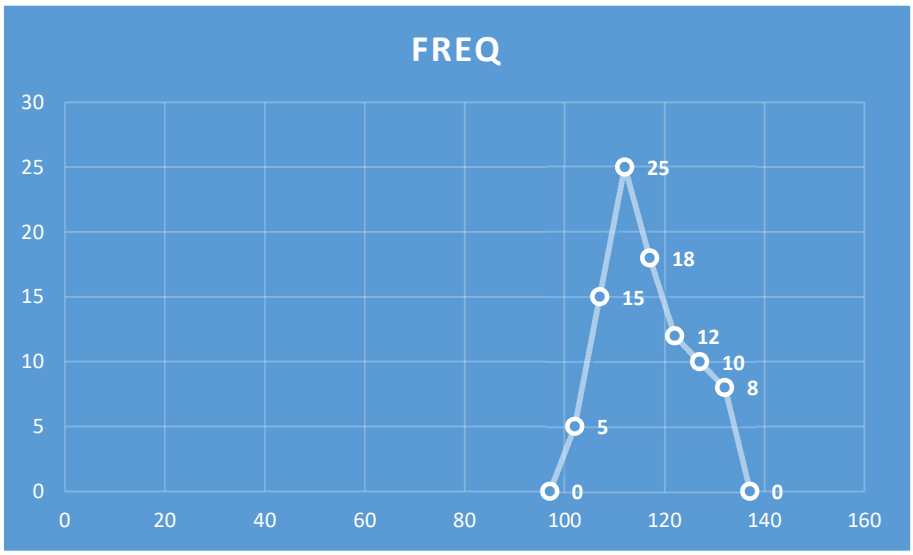
class limit	mid point	freq
100-104	102	5
105-109	107	15
110-114	112	25
115-119	117	18
120-124	127	12
125-129	132	10
130-134	137	8

Statistics – iGAP Technologies Pvt. Ltd



but lines for the freq polygon begin and end on x axis.
97 is difference

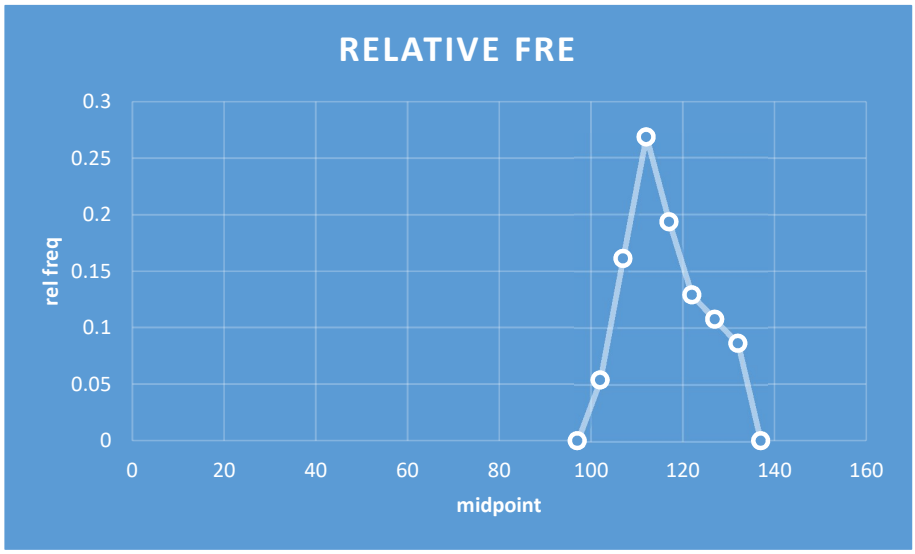
class limit	mid point	freq
	97	0
100-104	102	5
105-109	107	15
110-114	112	25
115-119	117	18
120-124	122	12
125-129	127	10
130-134	132	8
	137	0



Relative freq polygon

class limit	mid point	freq	relative fre
	97		0
100-104	102	5	0.05376344
105-109	107	15	0.16129032
110-114	112	25	0.2688172
115-119	117	18	0.19354839
120-124	122	12	0.12903226
125-129	127	10	0.10752688
130-134	132	8	0.08602151
	137	93	0

rel freq = freq/total



Lecture 16 - Biomodal and U shaped distribution

Biomodal –

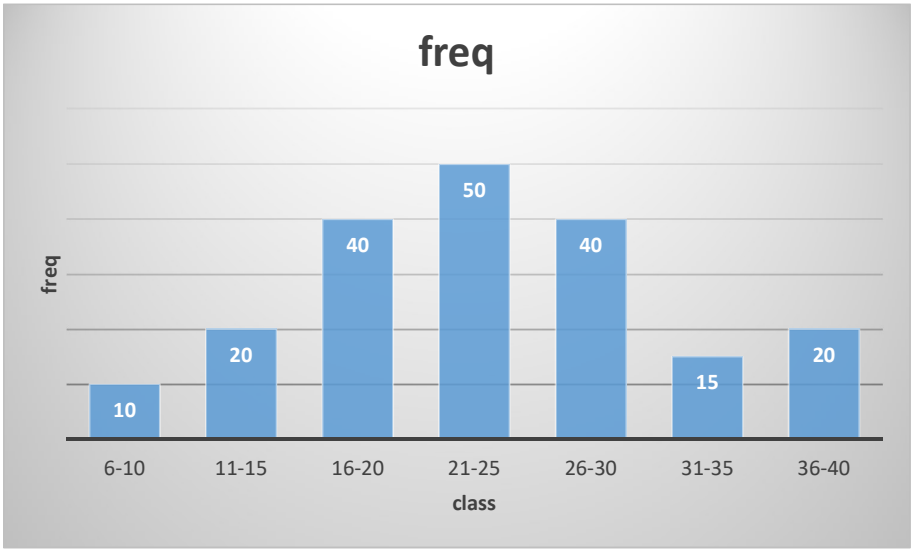
A data set is bimodal if it has two modes. This means that there is not a single data value that arises with the highest frequency.

If mode occurs 2 times, it is called bimodal.

Bimodal

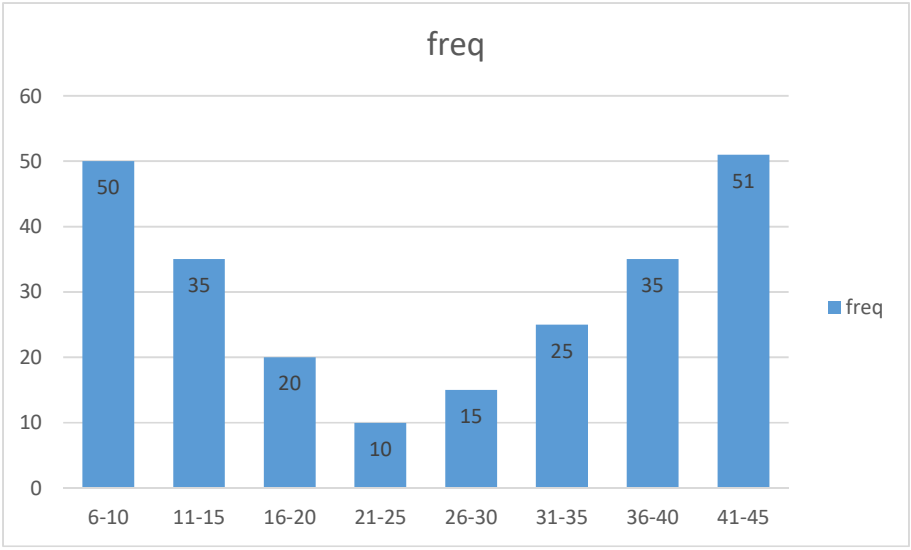
class	freq
6-10	10
11-15	20

16-20	40
21-25	50
26-30	40
31-35	15
36-40	20



U
shaped

class	freq
6-10	50
11-15	35
16-20	20
21-25	10
26-30	15
31-35	25
36-40	35
41-45	51



Statistics – iGAP Technologies Pvt. Ltd

Mean -

Learning Objectives: • Summarize data, using measures of central tendency, such as the

• Mean,

• Median,

• Mode,

• Midrange, and

Weighted Mean.

• Measures of average are called measures of central tendency.

Examples: The average height of women is 5 feet and 3 inches.

The average marks are 80%.

The average speed in the school zone is 30km/hr. etc.

•A statistic is a characteristic or measure obtained by using the data values from a sample.

• A parameter is a characteristic or measure obtained by using all the data values for a specific population.

• The Mean is the quotient of the sum of the values and the total number of values

• The symbol X is used for sample mean. (Roman letters are used for Statistics) X: X1 +X2 + X3 + x4 + Xn/n;

EX • For a population, the Greek letter u (mu) is used for the mean. • Greek letters are used for parameters. u=: X, + X, + X, +*4 +X, _ EX

Notebook – show

Lecture 18 – independent event vs dependent events

Independent

• Two events are independent if the outcome of the first event does not affect the outcome of the second event.

Ex : coin

Ex - •On the patio of Pizza Hut, the Sarah's family decides on a type of pizza to order.

In the back of the restaurant, another family also chooses a pizza to order. Are these two events dependent or independent?

•Independent

The two events are independent. The pizza that the Sarah's family orders does not affect which pizza another family orders.

Dependent –

Ashley purchases a new car from a dealership in New York. An hour later, another customer buys car at that same dealership. Are these two events dependent or independent! Dependent

• The two events are dependent. Ashley's choice a car affects which ones are left at the dealership for the second customer to buy.

Lecture 19 – Measures of central Tendency 2

Median –

• The median is the midpoint of the data array.) The symbol used for the median is MD.

• The raw data must be arranged in ascending order, and when the data is arranged, it is called data array.

• The median will be a specific value in the data set if the data is in odd numbers, or will fall between two values if the data set is even.

• Arrange the data values in ascending order or increasing order.

• determine the number of values in the data set. If numbers are odd, select the middle data value as the median. If numbers are even, find the mean of the two-middle values. (add them and divide by 2).

Find the median of the following data set.

A) 12,34,23,45,67 odd data set

• Arrange in order

12, 23, 34, 45, 67

• B) 23, 56,87,34, 12,76 12, 23, 34, 56, 76 even data set
12,23,34,56,76,87 data array

• Add 34 + 56 =96/2
MD 48 is the median.

Lectures 20 – Measures of central Tendency 3

Mode –

•The number which appears most often in a dataset of numbers is called mode.
•It is sometimes said to be the most important item.
• There may be no mode, one mode (unimodal).
two modes (bimodal), or many modes (multimodal).

• Find the Mode from the following data set.
• 10, 23,45, 67, 10, 10,67
• Arrange in order but necessary for mode.
10, 10, 10,23,45,67,67 ;
Mode is 10
The data set is said to be unimodal .
• Find the Mode from the following data set.
• 20,23,45,69,20,20,69,69
• 20,20, 20,23,45,69, 69,69 ;
Mode 20 and 69;
The data set is said to be bimodal.

• Find the Mode from the following data set.
• 24,61,30,69,20, 24,69,61, 45, 48,45,24,45,61 »

20,24,24,24,30, 45,45, 45,43,61,61,61 69,69

• Mode 24, 45 and 61 The data set is said to be multimode. Having more than two modes is called "multimodal.

Lectures 21 - Measures of central Tendency 4

Midrange –

• The arithmetic mean of the largest and the smallest values in a data set.
The midrange is the average of the lowest and highest values in a data set.

• Find the midrange of the following data set.
12,34,56,98, 10
= 10 +98 /2
= 108 /2
= 54 is midrange

Lectures 22 - Measures of central Tendency 5
The weighted Mean –
The Weighted Mean Means that is calculated with extra weight given to one or more elements of the sample. .
• The weighted mean is similar to an ordinary arithmetic mean (the most common type of average), except that instead of each of the data points contributing equally to the final average, some data points contribute more than others.
Example - notebook.

Lectures23 – Measures of variation

Learning Objectives: Spread or Variability of Data Set:
1. Range
2. Variance
3. Standard deviations

In order to describe the data set more accurately, statisticians use measure of variations instead of measure of central tendency.

• Measure of central tendency
1. Mean
2. Median
3. Mode
4. Midrange
5. Weighted Mean

• Measure of variations
1. Range
2. Variance
3. Standard deviations.

Example – light bulb. Notebook

Statistics – iGAP Technologies Pvt. Ltd

• Two different brands of light bulbs are tested to see how long each will last before it fuse. Seven bulbs of each brand make a small population. The results (in weeks) are shown. Find the mean and range of each group.

Class A	class B
20	10
34	20
30	30
35	40
45	50
40	49
55	60

Uses of the Variance and Standard Deviation
•To determine the spread of the data.
•To determine the consistency of a variable.
•To determine the number of data values that fall within a specified interval in a distribution.
•Used in inferential statistics.

Measures of Variation: Range

The range is the difference between the highest and lowest values in a data set, and is simplest of the three measures.
R = Highest - Lowest

Example: 20, 35, 67, 90, 75, 10 R
= 90 - 10
= 80

Comparison of the light bulbs.

Range of Brand A	Range of Brand B
• R = Highest - Lowest = 55 - 20 = 35	• R = Highest - Lowest = = 60-10 = 50

• It concludes that 35 weeks separate the largest data value from the lowest data value.

• It concludes that 50 weeks separate the data value from the lowest data value.
•To see the more meaningful statistic to measure the variability, statisticians use Variance and Standard Deviations.
•Population Variance and Standard Deviations
•Sample Variance and Standard Deviations

How to find the Population Variance and Population Standard Deviation.

Notebook –

Lecture24 - Measures of variation
Learning Objectives:
Sample Variance and Standard Deviation

Formula for sample Variance •
Formula
• X = Individual value
• X = sample mean
• n = sample size

Lecture 25 – variance and standard deviation
Notebook

Lecture 26 – variance and std deviation of grouped data

Variance and Standard deviation for grouped data
1. Find the mid point of each class/class boundaries
2. Multiply the frequency by the mid point for each class.
3. Multiply the frequency by the square of the mid point for each class.
4. Find the sum of column B, D, E. (note: the sum of column B represents n, the sum of column D represents summation of f*x,. the sum of column E represents summation of x*m square.
5. Put the values in the formula and calculate variance.
6. Take the square root for standard deviation.
Notebook –

Lecture 27 – coefficient of variation
• The coefficient of variation CV or (Var) is the ratio of the standard deviation to the mean.

• It allows for comparison between distributions of values whose scales of measurement are not comparable. It is represented by percentage.
Notebook –

Lecture 28 – range rule of thumb

• The range is approximately four times the standard deviation. The standard deviation is another measure of spread in statistics. It tells you how your data is clustered around the mean.
• Note: Range rule of thumb is only for rough estimation.

The Range Rule of Thumb approximates the standard deviation as, If the distribution is unimodal and approximately symmetric.

S = range/4

Note: Range rule of thumb is only for rough approximation, and should be used when the distribution of data values are unimodal and roughly symmetric.
Example - notebook.

The range rule of thumb can also be used to approximate the largest and smallest data value.
• The smallest and largest data value will be approx. 2 standard deviations below and above the mean respectively.
• Smallest data value = Xbar - 2s = 9 - 2(2.83) = 3.34
Largest data value = Xbar + 2s = 9 + 2(2.83) = 14.66
• Note: These are rough approximation, for many data sets, almost all within 2 standard deviations of the mean. We will see better approx. in Chebyshev's theorem and in Empirical rule.

Lecture 29 – Chebyshev's Theorem

• Chebyshev's Theorem is a fact that applies to all possible data sets. It describes the minimum proportion of the measurements that lie must within two, three or more standard deviations of the mean.

• The proportion of values from any data set that fall within k standard deviations of the mean will least 1 - 1/ksquare, where k is a number greater than 1 (k is not necessarily an integer).

1 - 1/ ksqr
1 - 1/ ksqr
= 1-1/2sqr
= 1-1/4
= 3/4
= 75% at least

= 1-1/3sqr
= 1-1/9
= 8/9
= 88.89% at least

This theorem states that at least three fourths, or 75% of the data values will fall within 2 standard deviations of the mean of the data, while eight ninths or 88.89% of the data values will fall within 3 standard deviations of the mean and so on.
Example – notebook

Lecture 30 – Empirical rule (Normal)

• The Empirical Rule applies to a normal, bell-shaped curve, so the results are more accurate than Chebyshev's theorem.
• It states that within one standard deviation of the mean (both left-side and right-side) there is about 68% of the data.
• within two standard deviations of the mean (both left side and right-side) there is about 95% of the data. and within three standard deviations of the mean (both left-side and right-side) there is about 99.7 % of the data.

The percentage of values from a data set that fall within k standard deviations of the mean in a normal (bell-shaped) distribution is listed below.

No of std deviation k	proportion of k std deviation
1	68%
2	95%
3	97%

Statistics – iGAP Technologies Pvt. Ltd

The empirical rule can be broken down into three parts: 68% of data falls within the first standard deviation from the mean. 95% fall within two standard deviations. 99.7% fall within three standard deviations.

Exp – notebook

Lecture 31 – z score or standard score

Learning Objectives:

Standard scores

Percentiles

Deciles

Measures of Position

• A measure of position is a method by which the position that a particular data value has within a given data set can be identified. Or

• They are used to locate the relative position of the data value in the data set. Or

• They are used to find the position of a value, relative to other values in a set of observations/data. The most common measures of position are percentiles, quartiles, Deciles, and standard scores (z-scores)

Standard Score or Z Score

• A z-score is also known as a standard score. Z-score are expressed in terms of standard deviations from the mean

• The z-score tells how many standard deviations a data value is above or below the mean.

• If a z-score is equal to 0, it is on the mean.

• If a Z-Score is equal to +1, it is 1 Standard Deviation above the mean.

• If a z-score is equal to +2, it is 2 Standard Deviations above the mean.

Formula for Z Score

• A z-score or standard score for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation.

Z = Value - mean / Standard Deviations

Statistics – iGAP Technologies Pvt. Ltd

• For Samples

z = x – xbar/s

For Populations

z = x-mu/sigma

Lecture 32 – how to calculate percentile

Percentiles

• A percentile is a value below which a certain percentage of observations lie.

• Percentile divide the data set into 100 equal groups/parts.

• Percentiles are position measures used mostly in educational and health related fields to indicate the position of an individual in a group.

1 Percentiles are not the same as percentages. .

• Example: If a student gets 67 points in a test out of 100, it means she has 67%. There is no indication of her position with respect her class. May be her score is highest, the lowest somewhere in between.

Example – notebook

Finding a data value corresponding to a given percentile.

1. Arrange the data in order from lowest to highest.

2. Put the value into the formula c = n* P /100

n = total number of values

p = percentile

3. If c is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded up value.

4. If c is a whole number, use the value half way between the cth and (c+1)st when counting up from the lowest value.

Example:

The marks of 10 students are given below. Find the value corresponding to 25th percentile.

76,56,59,87,90,34,49,48,75,62

Solution:

Arrange the data in order:

34, 48, 49, 56, 59, 62, 75, 76, 87, 90

Put the values in formula:

C = n*P /100

= 10 * 25 /100

C = 2.5

If c is not a whole number, round up to the next whole number. So

c = 3, now start at lowest value and count over to the third value, which is 49, so the value 49 corresponds to the 25^{th} percentile.

Example2:

The marks of 10 students are given below.

Find the value corresponding to 70th percentile.
76,56,59,87,90,34,49,48,78,62

Solution: Arrange the data in order:

34, 48, 49, 56, 59, 62, 76, 78, 87, 90

Put the values in formula:

c = n*P /100

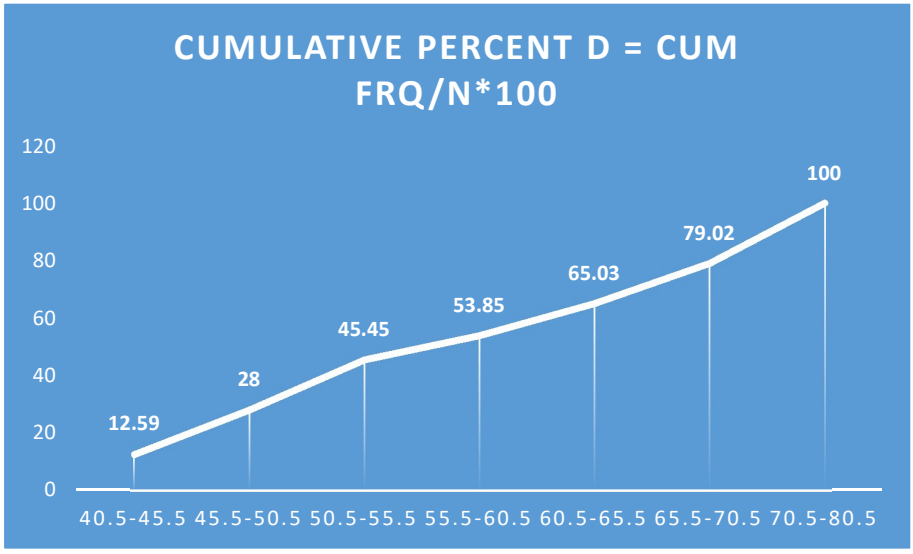
c = 10 * 70 /100

c = 7

Since c is a whole number, use the value halfway between the c and

(c +1) value when counting up from the lowest value.in this case 76 and 78 add them and divide by 2.it means 77 corresponds to 70^{th} percentile means obtaining 77 marks would have done better than 70% of the class.

class boundries	freq	cumulative freq	cumulative percent
a	b	c	d = cum frq/n*100
40.5-45.5	18	18	12.59
45.5-50.5	22	40	28
50.5-55.5	25	65	45.45
55.5-60.5	12	77	53.85
60.5-65.5	16	93	65.03
65.5-70.5	20	113	79.02
70.5-80.5	30	143	100
	n = 143		



Lecture 33 – Quartiles

Steps to find Quartiles:

1. Arrange the data in order from lowest to highest.

2. Find the median of the data values. This is the value for Q2-

3. Find the median of the data values that fall below Q2. This is the value for Q1•

4. Find the median of the data values that fall above Q2 This is the value for Q3.

*Interquartile Range IQ is the difference between the third and first quartile.

IQR = Q3 - Q1

1, 2, 3, 4, 4, 5, 6, 8, 9, 10, 12

IOR = 9 - 3

= 6

Note: In addition to dividing the data set into four group, quartiles can be used as rough measure of variability. This measure Of variability which uses quartiles is called the interquartile range and is the range of the middle 50 % of the data values. the more variable the data set is the larger the value of interquartile range will be.

Lecture 34 – Deciles

Deciles

• Decile divide the distribution into 10 groups, and is denoted by D1, D2 D3 and so on.

Formula D = k/10 * (n +1)

Example: Find Decile D7 and D8 of the following 10 students.
34,43,80,50,60,92,51,49,65,72

• Arrange the data in order:

34, 43, 49, 50, 51, 60, 65, 72, 80, 92

• Formula:

D = k/10 * (n +1)

D=7/10 * (10 +1)

D= 77/10

D = 7.7

D = 8 approx.

D7 is the 8th element, which is 72.

Find D8

Arrange the data in order:

34, 43, 49, 50, 51, 60, 65, 72, 80, 92

• Formula:

D = k/10 * (n +1)

D = 8/10 * (10 +1)

D = 88/10

D = 8.8

D = 9 approx.

D is the 9th element, which is 80.

Lecture 35 – Outlier

An outlier is an extremely high or low data value when compared with the rest of the data values.

Find the outliers from the following data set.

10, 20,30, 40,500

500 is outlier because extremely high value.

Find the outliers from the following data set.

10, 206, 240,300,350

10 is outlier because 10 is extremely low value.

A data value less than Q1 - 1.5(IQR) or

greater than Q3 + 1.5(IQR) can be considered an outlier.

Steps to find outliers

• Arrange the data in order from lowest to highest and find Q1 and Q3

• Find the interquartile range Q3 – Q1

• Multiply IQR by 1.5.

• Subtract step 3 from Q1 and add in Q3. Check the data set for any data value that is smaller than Q1- 1.5(IQR) or larger than Q3 + 1.5(IQR).

Example; find the outliers from the following data set

10, 11, 15,25,35,30,7,68 ..

Arrange the data in order from lowest to highest and find Q1 and Q3.
7,10,11,15,25,30,35,68

Q1 = 10.5

Q3 = 32.5

2.Find the interquartile range

IQR = Q3 - Q1.

= 32.5 - 10.5

= 22

3.Multiply IQR by 1.5

=33

4.Subtract IQR from Q1 and add in Q3

10.5 - 33= -22.5

32.5 + 33 = 65.5

5. check data set for any data value that is similar than Q1-1.5(IQR)

OR larger than Q3+1.5(IQR)

68 is the Outlier.

Lecture 36 – how to make boxplot

Exploratory data analysis

• In statistics, exploratory data analysis (EDA) is an approach to analysing data sets to summarize their main characteristics, frequently with visual methods.

•The purpose of exploratory data analysis is to study data to find out what information can be revealed about the data, such as centre (median) and the spread (Variation).

• In EDA data are represented graphically using a box plot or whisker plot.

Box Plot

• A box plot is a graphical version of statistical data based on the minimum value, maximum value, first quartile (Q1), third quartile (Q3),

and median. The "box plot" or graph looks like a rectangle with lines extending from the top and bottom.

Five points for Boxplots

1. Lowest value of the data set.

2. Highest value of the data set

3. Q1 Quartile one

4. Q3 Quartile three

5. The Median

How to make a Boxplot

•Write the data in order if it is not in order.

•Find the median of the data.

•The median divides the data into two halves. Find Q1, and Q3

•Find the lowest value

•Find the highest value

1. Draw a horizontal axis with a scale that includes the maximum and minimum data values.

2. Draw a box with vertical sides through Q1, and Q3, and draw a vertical line though the median.

3. Draw a line from the minimum data value to the left side of the box and a line from the maximum data value to the right side of the box.

•A boxplot can give information regarding the shape of the distribution, variability, and centre (or median) of a statistical data set.

•In a box and whisker plot, the ends of the box are the upper and lower quartiles, so the box spans the interquartile range.

Lecture 37 – sample space

Learning Objectives:

Determine sample spaces with observations and experiments.

Sample Space is the set of all possible outcomes of a probability experiment.

Experiment	Sample Space

Statistics – iGAP Technologies Pvt. Ltd

Student attendance	Present, Absent/leave
Toss one coin	Head, Tail
Toss 2 coins	H-H, H-T, T-T, T-H
True, False Questions	T, F

Sample Space using a Tree Diagram of three children in a family.

• A tree diagram display all the family possible outcomes of an event. Each branch in a tree diagram denotes a likely outcome.

• There are eight possibilities.

BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG

Lecture 38 – complementary events

• The Complement of an event is all outcomes that are not the event. Example:

For dice, when the event is (1, 5) the complement is (2, 3, 4,6). Together the event and its complement make all possible outcomes, or sample space.

• or The Complement of an event E, denoted by Ebar, is outcomes in the sample space that are not included in the outcomes of event E.

Find the Complement of the following

Events	Complement of the events
• Rolling a die and getting a 5.	• Getting a 1, 2, 3, 4, or 6.
• Selecting a month that has 28 days	• Jan, & Mar to Dec

That is feb

Together the event and its complement make all possible outcomes, or sample space. The sum of the probability of the event and the probability of its complement will equal to 1.

P(E) + P(Ebar) = 1

If 2 coins are tossed, so their sample space is HH, TT, HT, TH. If event E is all TT, which will be 1/4.

Its complement will be 3/4.

Put this value in above formula

P(E) + P(Ebar) = 1

1/4 + 3/4 =1

4/4=1

1=1.

Rules for Complementary Events.

• P(E) + P(Ebar) = 1

P(Ebar) = 1- P(E)

• P(E) = 1 - P(Ebar)

• Note: If the probability of an event or the probability of its complement is known, then the other can be found by subtracting the probability from 1.

Lecture 39 – not mutually exclusive event

Formula for not mutually exclusive event

• If two events suppose A and B are not mutually exclusive then:

P (A or B) = P(A) + P (B) - P (A and B)

• Example: A Single card is drawn from an ordinary deck of cards. Find the probability that it is either an ace or a red card.

• Solution:

P (A or B) = P(A) + P (B) - P (A and B)

= 4/52 + 26/52 – 2/52

= 28/52

=7/13

= 0.538

Explanation: Since there are 4 aces and 26 red cards (13 spades and 13 clubs), 2 of aces are red cards (spade and club). So the probability of two outcomes must be subtracted since they have been counted twice.

Example 2:

Statistics – iGAP Technologies Pvt. Ltd

In a School there are 10 teachers and 6 educational assistant. Out of them, there are 7 female teachers, and 4 female educational assistants. If a staff person is selected, what is the probability that the person selected is a teacher or male.

staff	female	male	total
Teachers	7	3	10
Educational assistance	4	2	6
total	11	5	16

Solution:

P (A or B) = P(A) + P (B) - P (A and B)

P(teacher or male) =P(teacher)+P(male)-P(teacher male)

= 10/16 + 5/16 -3/16 =12/16 P(Teacher or male)

= 0.75

Lecture 40 – probability rules

• The probability of any event E is a number (either a fraction or decimal) between and including 0 and 1.

This is denoted by $0 \leq P(E) \leq 1$.

• The sum of the probabilities of all the outcomes in the sample space is 1.

• If an event E cannot occur (i.e., the event contains no members in the sample space), its probability 0.

If an event E Is certain then the probability of e is 1.

Rule 1:

it says, the probability range is from 0 to 1. When the probability of an event is close to zero, its occurrence is unlikely. When the probability of an event is 0.5, there is about 50% chance that event will occur, and when the probability of an event is close to 1, the event is highly likely to occur.

• Rule 2:

In rolling a single die, each outcome in the sample space has a probability of 1/6, and the total of all outcomes equal to 1.

Rule 3:

When a single die is rolled, what is the probability of getting a number greater than 6? Since all sample space: 1, 2, 3, 4, 5, and 6, it is impossible having greater than 6, so the probability is O impossible having greater than 6, so the probability is 0

• Rule 4: When a single die is rolled, what is the probability of getting a number less than 8?

Since all outcomes: 1, 2, 3, 4, 5, and 6 are less than 8 the probability is 1.

P(number less than 8) = 6/6 = 1.

The event of getting a number less than 8 is certain.

Lecture 41 – classical probability

Classical probability

• Classical probability is the statistical model that measures the possibility of something happening, but in a classic sense. It also means that every statistical experiments contain elements that are equally likely happen.

All outcomes in the sample space be equally likely to occur.

• Example:

Classical probability is a simple form of probability that has equal chances of something happening. For example: Rolling a die. It's equally likely would get a 1, 2, 3, 4, 5, or 6. or each outcome has a probability of 1/6.

• When a card is selected from a deck, has a same probability of being selected that is 1/52 .

A sample space is the set of all possible outcomes of a probability experiment. A probability experiment is a chance process that leads to well defined results called outcomes.

• An event consists of outcomes of the probability experiment.

• An event can be one outcome or more than one outcome.

Example. if a die is rolled and a 4 comes up, this is called an outcome. An event with one outcome is called simple event. The event of getting

an even number when a die is rolled is called a compound event. it has three outcomes.(2,4,6).

• An outcome is the result of a single trial of a probability experiment. It consists of two or more outcomes.

Formula for classical probability

The probability of any event E is

= Number of outcomes in E / Total number of outcomes in the sample space

This probability is denoted by

P(E) = n(E) / n(S)

n(E) = no of outcome in event.

n(S) = no of outcome in sample space.

Note – probability can be expressed as a fraction, decimal or percentage.

Probability of getting head can be expressed 1/2 or 0.5 or 50%.

Example:

Find the probability of getting a black face card, (jac queen, or king) Total cards are 52, and 6 are black (jack, queen, or king) diamond and hearts. Number of outcomes in E Total number of outcomes in the sample space This probability is denoted by

P(E) = n(E) /n(S)

= 6/52

= 3/26

= 1/13

Find the probability that two of the three kids are girls, if a family has 3 kids. Sample Space: BBB, BBG BGB GBB GGG GGB GBG BGG

Three outcomes (GGB, GBG, BGG) have two girls.

The probability of having two of three children being girls is 3/8.

Find the probability of getting (1) A queen, (2) A spade (3) A black card, from an ordinary deck.

•P(queen) = 4/52 =1/13

•P(spade) = 13/52 = 1/4

P(black Card) = 26/52

= 1/2.

Lecture 42 – Empirical probability

Empirical probability

• Empirical probability is probability based on data collected through an experiment or observation. These probabilities are found by dividing the number of times an event occurred in an experiment by the total number of trials or observations

• Empirical probability relies on actual experience likelihood of outcomes

Formula for Empirical Probability

p(E) = Frequency for the class / Total frequencies in the distribution

= f/n.

Example:

Kidney transplant patients stayed in hospital for the following number of days.

Hospital days	frequency
5	12
8	20
12	15
7	7
6	13
9	16
	Total 83

P(E) = f/n

P(5) = 12/83

P(8) = 20/83

P(12) = 15/83

P(7) = 7/13

P(6) = 13/83

P(9) = 16/83

1. Patient stayed less than 9 days.

=12/83 +20/83 +7/83 +13/83 .

2 .Patient stayed at most 6 days

= 12/83 + 13/83

3. Patients stayed at least 12 days. –

= 15/83

Example2 - blood type of people

p(E) = f / n

Blood type	frequency
AB	12
O	20
A	15
B	7
total	54

1. People have AB
   = 12/54
2. People have O
   = 20/5
3. people have A

   = 15/54

4. people have B
   = 7/54
5. People have AB or 0.
   P(AB or 0) =12/54 + 20/54
6 People have 0 or A. P(A or 0)
   =20/54 +15/54

7. People have neither A nor B.

P(neither A nor B)

=12/54 +20/54 4.

People have not B.

P(not B) = 1 - P(B) = 1-7/54

= 47/54 .

Lecture 43 – counting rule

In order to find out the all possible outcomes for the sequence of events, three rules can be used.

1. Fundamental counting rule.

2. Permutation rule

3.Combination rule.

Fundamental counting rule:

• In a sequence of n events in which the first one has k1 possibilities and second event has k2 and so forth, then the total number of possibilities of the sequence will be

K1 * k2 * k3 * k4 ... kn

Note – the fundamental counting rule is also called the multiplication of choices.

Example: Pick two letters suppose A and B, and rolling a die. Find the total number of outcomes for the sequence of an events.

As there are 2 events A and B, and six (6), outcomes of rolling die, i.e. 1,2,3,4,5 and 6.

2 *6 = 12

there are 12 possibilities. A tree diagram can also be drawn for the sequence of an events.

Example:

There are 2 major routes from Toronto to New York, and 3 major routes from New York to California. How many different trips can be made from Toronto to California?

Solution: 1

2*3 = 6

6 different trips can be made from Toronto to California

Statistics – iGAP Technologies Pvt. Ltd

Tree diagram

Notebook –

Lecture 44 – subjective probability

• Subjective probability is a probability resulting from an individual's own judgment about whether a particular outcome is likely to occur.

It contains no formal calculations and only returns the subject's opinions and past experience.

• They differ from person to person, and because they are subjective, they can be formed on a person's opinions or other elements.

• Or

• Subjective probability uses a probability value based on an educated guess or approximation, employing views and inexact information.

Examples: weather forecasting, prediction of sports outcomes, etc.

• Examples:

• 1.The sportsman may say that there is 80% probability that their team will win the Olympics.

• 2. A doctor may say, on the basis of his diagnosis, that there is 50 % chance of kidney transplantation of the patient.

• 3. Agriculturists give opinion on the basis of their experience that there it 60% probability, that some crops will not good this year.

Lecture 45 – permutation Rule

Permutation rule:

• Permutation is an arrangement of n objects in a specific order,

• Permutation uses factorial notation. The factorial notation use exclamation mark.

4! = 4 * 3 * 2 * 1

8! = 8 * 7 * 6 * 5 * 4 * 3 * 2 * 1

Factorial is the product of all the positive numbers from 1, to a number.

Example: A home buyer has a choice of 6 houses to buy. He decides to rank each house according to certain criteria such as price of the house

and location. How many different ways can he rank the 4 houses? As there are 6 choices

6! = 6 * 5 * 4 * 3 * 2 * 1 = 720

There are 720 different ways.

Suppose he wishes to rank only the top 3 of the houses.

How many different ways can he rank them.

6*5*4 = 120.

Permutation Rule

1: Find the number of ways that "r" objects can be selected from "n" objects.

The arrangement of an n objects in a specific order using r objects at a time is called permutation of n objects taking r objects at a time.

It is written as nPr, and formula is

nPr = n! / (n - r)!

= 6! / (6 - 6)!

= 6*5*4*3*2*1 / 0!

= 720 different way: (0! = 1)

Suppose he wishes to rank only the top 3 of the houses. How many different ways can he rank them?

6P3 = 6! / (6 -3)!

= 6*5°4*3*2*1 / 3!

= 6*5*4*3*2*1 / 3*2*1

= 120 different ways

Permutation Rule 2:

The number of permutations of n objects when r1 objects are identical, r2 object are identical... rp objects are identical, etc.

n!/r1!r2!..rp!

Where r1 +r2+r3 ....rp = n ....

How many permutations of the letters can be made from the word ACCOUNTING.

In word Accounting. There are 1A, 2 C's 10, 1 U, 2 N, 1 T, 1I, 1 G

= n! / r1!r2!..rp!

= 10!/ 1*2!*1*1*2!*1*1*1

= 10*9*8*7*6*5*4*3*1 / 1*2*1*1*1*2*1*1*1*1

= 907,200 permutation that can be made from the word accounting.

Lecture 46 – counting rule - combination

A selection of distinct objects without regard to order is called a combination.

• Order of selection is not important.

• Combination are used when the order or arrangement is not important, as in the selecting process. Suppose 10 students are to be selected from 100 students, as a member of student's council. 10 students represent a combination, as it does not matter who is selected first, second, third and so on.

Combination rule:

The number of combinations of r obiects selected from n obiects is denoted by nCr, and is given by the formula

nCr = n! / (n - r)! r!

• How many combinations of 5 objects are there, taken 3 at a time?

• Solution: nCr = n! / (n - r)!r!

= 5! / (5-3)!3!

= 5*4*3*2*1 / 2*1*3*2*1

= 120/12

= 10 combination

Difference between combination and permutation

• The difference between a combination and a permutation can be shown using the letters A, B, C, and D.

The permutations for the letters A, B, C, D are

AB BA CA DA

AC BC CB DB

AD BD CD DC

• In permutations, AB is different from BA. But in combinations, AB is the same as BA since the order does not matter in combinations. Therefore, if duplicates are removed from a list of permutations, what is left is a list of combinations, as shown.

AB BA CA DA

AC BC CB DB

AD BD CD DC

Hence, the combinations of A, B, C, and D are AB, AC, AD, BC, BD, and CD. (Alternatively BA could be listed and AB crossed out, etc.) The combinations have been listed alphabetically for convenience but is not a requirement.

Lecture 47 – permutation using Microsoft excel

Permutation rule

• Permutation is an arrangement of n objects in a specific order

• Permutation uses factorial notation. The factorial notation uses exclamation mark.

i.e. 4! =4*3*2*1

8! = 8 * 7* 6 * 5 * 4 * 3 * 2 * 1

Factorial is the product of all the positive numbers from 1 to number.

Example: A home buyer has a choice of 6 houses to buy. He decides to rank each house according to certain criteria such as price of the house and location. How many different ways can he rank the 4 houses? As there are 6 choices

6! = 6 * 5 * 4 * 3 * 2 * 1 = 720

There are 720 different ways.

Suppose he wishes to rank only the top 3 of the houses.

How many different ways can he rank them.

6*5*4 = 120.

Statistics – iGAP Technologies Pvt. Ltd

Permutation Rule

1: Find the number of ways that "r" objects can be selected from "n" objects.

The arrangement of an n objects in a specific order using r objects at a time is called permutation of n objects taking r objects at a time.

It is written as nPr, and formula is

nPr = n! / (n - r)!

= 6! / (6 - 6)!

= 6*5*4*3*2*1 / 0!

= 720 different way: (0! = 1)

Suppose he wishes to rank only the top 3 of the houses. How many different ways can he rank them?

6P3 = 6! / (6 -3)!

= 6*5°4*3*2*1 / 3!

= 6*5*4*3*2*1 / 3*2*1

= 120 different ways

Using Microsoft excel - excel file

Lecture 48 – how to calculate combination using Ms excel

A selection of distinct objects without regard to order is called a combination.

• Order of selection is not important.

• Combination are used when the order or arrangement is not important, as in the selecting process. Suppose 10 students are to be selected from 100 students, as a member of student's council. 10 students represent a combination, as it does not matter who is selected first, second, third and so on.

Combination rule:

The number of combinations of r obiects selected from n obiects is denoted by nCr, and is given by the formula

nCr = n! / (n - r)! r!

• How many combinations of 5 objects are there, taken 3 at a time?

• Solution: nCr = n! / (n - r)!r!

= 5! / (5-3)!3!

= 5*4*3*2*1 / 2*1*3*2*1

= 120/12

= 10 combination

Using Ms excel – excel sheet

Lecture 49 – how to use factorial using Ms excel

The factorial notation uses exclamation mark.

i.e.

4! =4*3*2*1

8! =8*7*6*5*4*3*2*1

Factorial is the product of all the positive numbers from 1 to a number.

Calculate – excel sheet

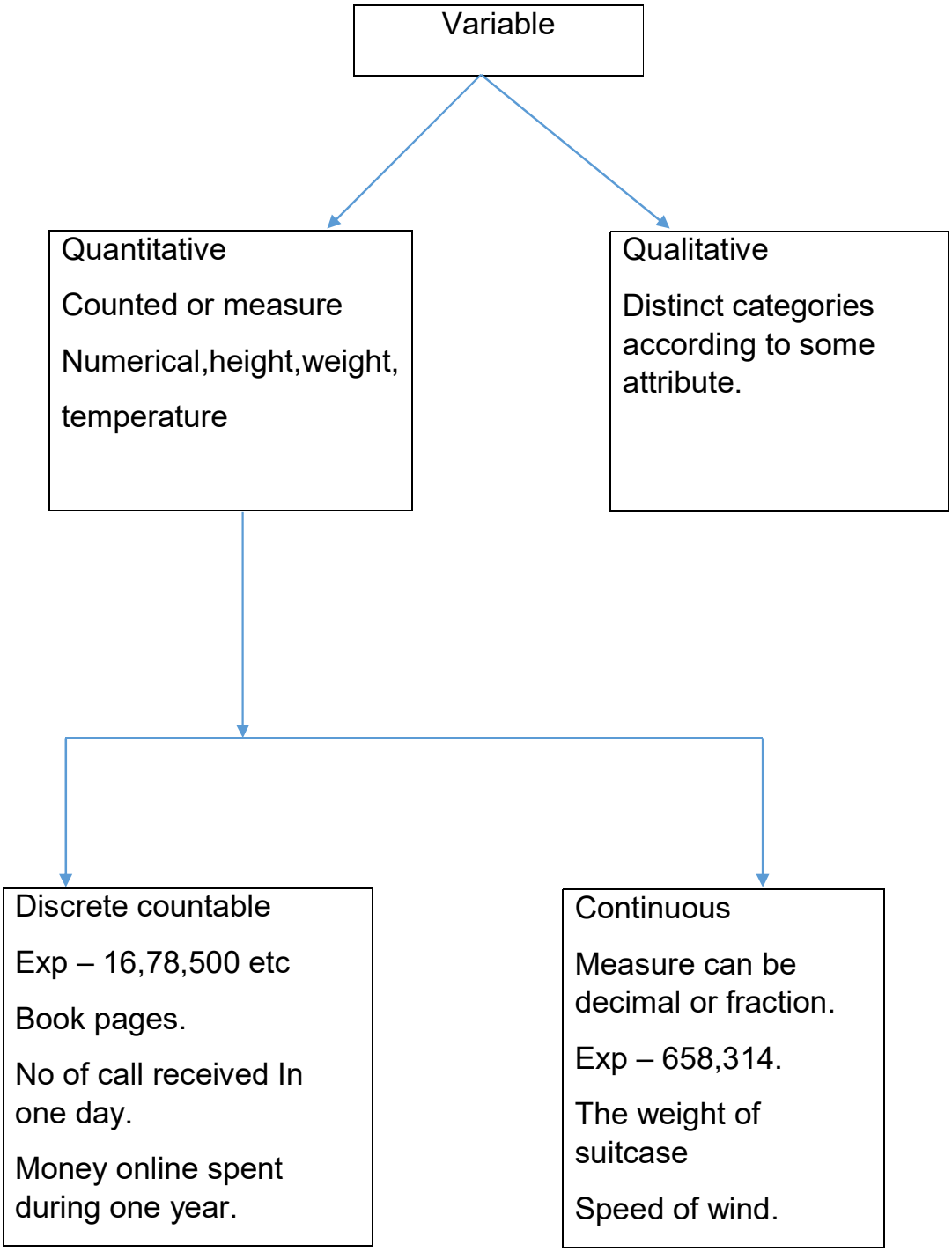
Lecture 50 – probability distribution

A variable is a characteristic or attribute that can assume different values. any letter can be used to represent the variables. X,Y,Z.

• Variables whose values are determined by chance are called random Variables.

• Example. Automobile insurance, claim supposed 5%every year.

Statistics – iGAP Technologies Pvt. Ltd



Discrete Probability Distribution

• A discrete probability distribution consists of the values a random variable can assume and the corresponding probabilities of the values.

Statistics – iGAP Technologies Pvt. Ltd

Example: Construct a probability distribution for a discrete random variables of the following sample space. BBB, BBG BGB GBB GGG GGB GBG BGG.

Solution. If X is the random variable for the number of girls, then it assumes the value, 0,1,2,3

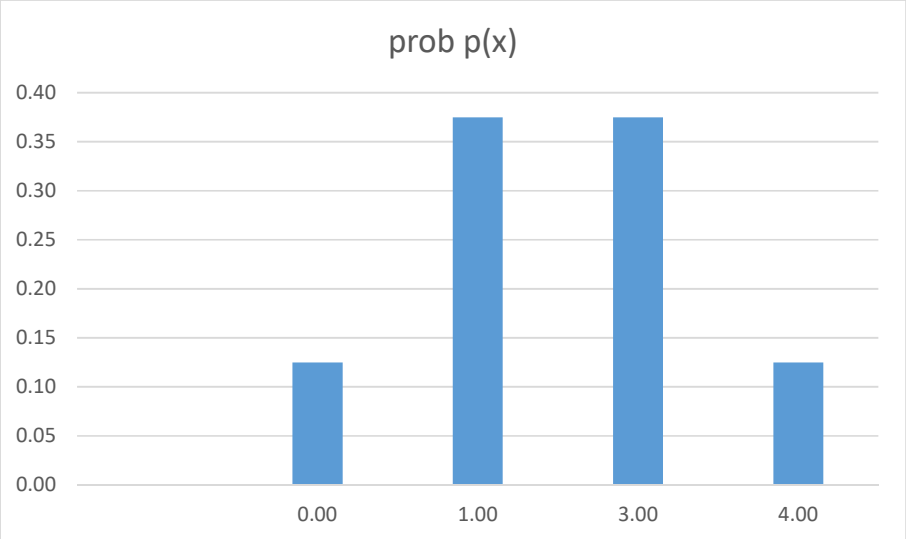
No girl	1 girl	2 girl	3 girl
BBB 1/8	BBG BGB GBB 3/8	GGB GBG BGG 3/8	GGG 1/8

Number of girls x	0	1	2	3
Prob p(x)	1/8	3/8	3/8	1/8

Graphical representation of Probability distribution.

• When probability distributions are shown graphically, the value of X are placed on x axis, and while P(X) are taken on Y axis. These graphs helps to find out the shape of the distribution, i.e. right skewed, left skewed or symmetric.

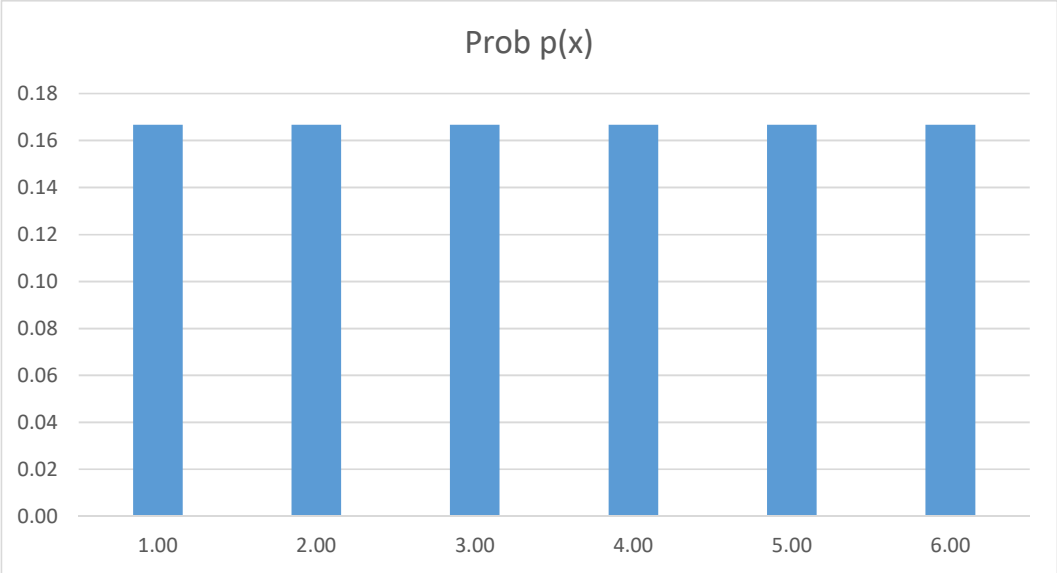
number of girl x	0.00	1.00	3.00	4.00
prob p(x)	0.13	0.38	0.38	0.13



Statistics – iGAP Technologies Pvt. Ltd

Example:
Construct a probability distribution for rolling a single die
• As the sample space of the die is 1,2,3,4,5,6 and each outcome has a probability of 1/6. The distribution is as follows

Outcome x	1.00	2.00	3.00	4.00	5.00	6.00
Prob p(x)	0.17	0.17	0.17	0.17	0.17	0.17



Two requirements for a probability distribution
1. The probability of any event E is a number (either a fraction or decimal) between and including 0 and 1.
2. This is denoted by $0 < P(E) < 1$.
Rule 1: it says, the probability range is from 0 to 1. When the probability of an event is close to zero, its occurrence is unlikely. When the probability of an event is 0.5, there is about 50% chance that event will occur, and when the probability of an event is close to 1, the event is highly likely to occur.

Rule 2 – in rolling a single die in outcome in sample space has probability of 1/6 and total of all outcome equal to 1.

•Example: The data shown consists of 45 world series events of basketball. Y represents number of games played in each series. Find the probability of P(Y) for each Y, construct a probability distribution

Solution: The probability of P(Y) is computed for each Y by dividing the number of series played for each Y by the total.

y	P(y)
5	12
3	14
6	9
4	10
total	45

————	6	3	6	4
Prob p(y)	0.2667	0.3111	0.2	0.222

= 12/45 = 0.2667
= 14/45 = 0.3111
= 9/45 = 0.2
= 10/45 = 0.222
Probability range 0 to 1

Find whether each distribution is a probability distribution

X	3	5	7	8
P(x)	0.3	0.5	0.3	0.1

y	4	6	4	6
p(y)	0.3	0.2	0.1	0.4

z	1	2	3	4
p(z)	-0.5	0.6	0.1	0.3

1. No, as the sum of probabilities is greater than 1

2. Yes, as the sum of probabilities is equal to 1.

3. No, as one of the probability is less than zero.

Lecture 51 – mean,variance,std deviation for prob distribution.

• Mean, variance, & standard deviation for a probability distribution are computed differently from the mean, variance,& standard deviation for samples or populations.

Formula used for the mean of probability distribution. The mean of a random variable with a discrete probability distribution is

$\mu = X1(PX1) + X2(PX2) + X3\ (PX3) + ...xn\ P(Xn)$

$= \sum X.\ P(X)$

X1, X2, ...Xn, are outcomes while P(X1), P (X2) ... are corresponding probabilities.

Example 1: find the mean of all the outcomes when the die is rolled.

Sol –

Example 2 : Find the mean of all the outcomes when three coins are tossed.

Formula used for the Variance and Standard deviation probability distribution.

$(sigma)\delta\ 2 = \sum x2*[p(x)] - \mu2$

$\delta = \sqrt{\sum} [x2*p[x] - \mu2$

Steps to compute Variance and Standard Deviation of Probability Distribution.

Statistics – iGAP Technologies Pvt. Ltd

1.Find the mean of all the outcomes.

2. Square each of the outcome.

3. Multiply the square of each outcome by its corresponding probability
4. Sum the products, and

5.Subtract the square of the mean from step 4.

6.Take the square root to find the standard deviation.

Example: Calculate the variance and standards deviation fi the probability distribution of rolling a die.

Lecture 52 – expected value of discrete random variable

"The expected value, of a discrete random variable of a probability distribution is the theoretical average of the variable.

•The expected value, or expectation, is by definition, the mean of the probability distribution.

Formula for Expected Value

$\mu = E(X) = \sum X.P(X)$

Example: 5 boxes are numbered like 3,5,7,8,9. A box is selected at random and its number is recorded and then it is replaced. Find the expected value of the number that will occur.

Number y	3	5	7	8	9
Prob p(y)	1/5	1/5	1/5	1/5	1/5

$\mu = E(y) = \sum x.p(x)$

$= 3*(1/5) + 5*(1/5) + 7*(1/5) + 8*(1/5) + 9*(1/5)$

$= 6.4$

Example: 1000 tickets were sold at $1 each for the washing machine valued at $450. What is the expected value of the gain if you purchase one ticket!

Gain z	win	lose
	$449	-1
Prob p(z)	1/1000	999/1000

1. For a win, the net gain is $449,as you will not get back$1 .

2. For a lose, gain is represented by negative number -1.

E(Z) = $449 * 1/1000 + (-1) * 999/1000

= 0.449 - 0.999

=$0.55

It means that a person would lose $0.55 average on each ticket purchased.

Lecture 53 – Binomial distribution

Binomial Experiment Many types of probability problems have only two possible outcomes or they can be reduced to two outcomes.

Examples: When a coin is tossed it can be heads or tails, True and False questions,

when a baby is born it is either a boy or girl,

In any game, you may win or lose etc. This kinds of situations are called binomial experiment.

Each repetition of experiment is called trial.

The binomial experiment is a probability experiment that satisfies these requirements:

1. There must be a fixed number of trials.

2. Each trial can have only two outcomes.

3. The outcomes of each trial must be independent of each other.

4. The probability of success must remain the same for each trial.

Examples: Interpret the situations as a binomial experiment.

1. Tossing a coin 50 times to see how many tails occur.

Yes, all four requirements are met.

2. Selecting 10 students from the class and recording their gender?

Yes, all four requirements are met.

3. Drawing 4 cards from a deck without replacement and recording whether they are black or red cards.

No, as the cards are not replaced, the events are not independent.

4. Asking a 50 people which brand of drink they like.

No as more brands are there.

Binomial distribution

• The outcomes of a binomial experiment and the corresponding probabilities of these outcomes are called a binomial distribution.

• In a binomial experiment, the outcomes are classified as a successes or failures, and are considered random variables. Its probability distribution is called the binomial distribution

• i.e. In MCQs correct answer is success, and all others are

Notation for the Binomial Distribution

P(S),	Probability of success
P(F)	X Probability of failure
p	The numerical probability of success
q	The numerical probability of failure
p(s)=p	and P(F) = 1 - p = q
n	The number of trials
x	The number of successes in n trials.

Note that x = 0,1,2,3…..n.

Binomial Probability Formula

In a Binomial experiment, the probability of exactly X success in n trials is computed by this formula.

P(X) = n! / (n - X)! X! * p raise to x * q raise to (n-x)

Find the probability that two of the three kids are girls, if a family has 3 kids. Sample Space: BBB, BBG BGB GBB GGG GGB GBG BGG Three outcomes (GGB, GBG, BGG) have two girls. The probability of having two of three being girl is 3/8

= 0.375

Solve the same problem by using the binomial formula We have to see whether it meets the four requirements.

1. There are fixed number of trails. 3 kids

2. There are only 2 outcomes for each trail Boy of Girl.

3. The outcomes are independent of one another.

4. The probability of success in each case is ½.

• In this case n = 3, X = 2, p = ½ and q = ½,

put the values in formula.

P(X) = n! / (n - X)! X! * p raise x * q raise (n-x)

= 3! / (3-2)!2! (1/2)2 * (1/2)1

= 3/8

= 0.375

This is same answer is obtained using the sample space.

Example: Suppose out of 5 people say, they likes Pepsi. If 10 people are selected at random, what is the probability that exactly 3 people like Pepsi.

Notebook – solve.

Mean Variance and Standard Deviation for Binomial Distribution can be found by using the following formulas.

• Mean μ = n. P

• Variance δ2 = n. P. q.

• Standard Deviation δ = √n*p*q

Example: Find the Mean, Variance and Standard deviation of the number of tails, if the coin is tossed 8 times.

• Mean μ = n. p

= 8. ½ Mean

= 4

Variance δ2 = n*p*q =

= 8. ½.1/2

= 2

Standard Deviation

δ = √n*p*q

= √2

= 1.41 approx.

Example: A die is rolled 750 times. Find the mean, variance and standard deviation of number 5 that will be rolled. Using binomial distribution.

Mean μ = n. P = 750 * 1/6

= 125

Variance δ2 = n. P. q

= 750 * 1/6 * 5/6 q = 1-p

= 104.16

Standard deviation δ2 = √n*p*q

= √104.16

= 10.206

Lecture 54 – normal distribution

• In probability theory, the normal distribution is a very common continuous probability distribution.

• Many continuous variables have distributions that are bell-shaped and are called approximately normally distributed variables.

• The theoretical curve, called the bell curve or the Gaussian distribution, (German Mathematician) can be used to study many variables that are not normally distributed but are approximately normal.

If the random variable has a probability distribution whose graph is continuous., bell shaped, and symmetric, it is called normal distribution. The graph is called the normal distribution curve.

Mathematical equation for normal distribution

$$y = \frac{e^{(x-\mu)2/2\delta^2}}{\delta\sqrt{2\pi}}$$

E = 2.718

Π = 3.14

μ = population mean

δ = population std deviation

The shape and position of the normal distribution curve depend on two parameters,

1. Mean and

2. Standard deviation.

Each normally distributed variable has its own normal distribution curve, which depends on the values of the variable's mean and standard deviation.

Properties of normal distribution

The normal distribution curve is bell-shaped.

The mean, median, and mode are equal and located at the center of the distribution.

The normal distribution curve is unimodal (only one mode). The curve is symmetrical about the mean, which is equivalent to saying that its shape is the same on both sides of a vertical line passing through the center.

The curve is continuous i.e., there are no gaps or holes. For each value of X, there is corresponding value of Y.

The curve never touches the x-axis. Theoretically, no matter how far in either direction the curve extends, it never meets the x-axis-but gets increasingly close.

The total area under the normal distribution curve is equal to 1.00 or 100%.

The area under the normal curve that lies within ----

one standard deviation of the mean is approximately 0.68 (68%).

two standard deviations of the mean is approximately 0.95 (95%)

three standard deviations of the mean is approximately 0.997 (97%)

Statistics – iGAP Technologies Pvt. Ltd