STATISTICAL METHODS 1

MATH 153

GYAMERAH, Samuel Asante (Ph.D.)[1]

[1]Department of Statistics and Actuarial Science
Kwame Nkrumah University of Science and Technology

2021/2022

Department of Statistics and Actuarial Science, KNUST

COURSE OUTLINE

1. Introduction to Statistics
2. Frequency Distributions and Graphs
3. Measures of Central Tendency
4. Measures of Variation
5. Measures of Position
6. Probability and Counting Rules
7. Random Variables
8. Discrete Probability Distributions

RECOMMENDED TEXT:

1. ELEMENTARY STATISTICS (A Step by Step Approach) by ALLAN G. BLUMAN

INTRODUCTION TO STATISTICS

- **Statistics** is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data

- Students study statistics for **several reasons** ;
  - Like professional people, you must be able to read and understand the various statistical studies performed in your fields. To have this understanding, you must be knowledgeable about the vocabulary, symbols, concepts, and statistical procedures used in these studies.

INTRODUCTION TO STATISTICS

- You may be called on to conduct research in your field, since statistical procedures are basic to research. To accomplish this, you must be able to design experiments; collect, organize, analyze, and summarize data; and possibly make reliable predictions or forecasts for future use. You must also be able to communicate the results of the study in your own words.
- . You can also use the knowledge gained from studying statistics to become better consumers and citizens. For example, you can make intelligent decisions about what products to purchase based on consumer studies, about government spending based on utilization studies, and so on

INTRODUCTION TO STATISTICS

1. **Variable** – characteristic or attribute that can assume different values
2. **Data** – consists of information coming from observations, measurement or responses. Data are the values (measurements or observations) that the variables can assume

3. Variables whose values are determined by chance are called **random variables**
4. A collection of data values forms a **data set**. Each value in the data set is called a **data value or a datum**
5. **Population** – the collection of **all** outcomes, responses, measurements or counts that are of interest.

6. **Sample** – a subset of a population, i.e a group of subjects selected from a population

QUESTIONS

1. In a recent survey, 2500 adults in Ghana were asked if they thought there were solid evidence for global warning. 1500 of the adults said yes. Identify the population and the sample

2. **Answer** : The population consists of the responses of all adults in Ghana.

3. **Answer** : The sample consists of the responses of the 2500 adults in Ghana in the the survey

Try

A survey of 500 freshmen in KNUST found that 95% did not get their first choice programmes. Identify the population and the sample

INTRODUCTION TO STATISTICS

1. **Parameter** – a number that describes a population characteristic.
   Average age of all people in Ghana.

2. **Statistic** – a number that describes a sample characteristic.

   Average age of people from a sample of three regions in Ghana.

   Decide whether the numerical value represents a parameter or a statistic.

   - A recent survey of a sample of college career centres reported that average starting for petroleum engineers is $\$83,121$
   - The average cut-off point for the 2182 students who accepted admission offers to KNUST in 2009 was aggregate 12

INTRODUCTION TO STATISTICS

- The average of $\$83,121$ is a statistic since it is based on the subset of the population.

Try

A survey of 500 adults in Ghana found that 54% take in beer daily. 54% is a parameter. TRUE or FALSE?

Branches of Statistics

- **Descriptive statistics** consists of the collection, organization, summarization, and presentation of data. In descriptive statistics the statistician tries to describe a situation.

- Example: Consider the national census conducted by Ghana Statistical Service (GSS) every 10 years. Results of this census give you the average age, income, and other characteristics of the population in Ghana. To obtain this information, GSS must have some means to collect relevant data. Once data are collected, GSS must organize and summarize them. Finally, GSS needs a means of presenting the data in some meaningful form, such as charts, graphs, or tables.

- **Inferential Statistics** – involves using a sample to draw conclusion about a population. Inferential statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions. Inferential statistics uses probability, i.e., the chance of an event occurring

- **Question**: A large sample of men, aged 48, was studied for 18 years. For unmarried men, approximately 70% were alive at age 65. For the married men, 90% were alive at age 65. Decide which part of this study represents the descriptive branch of statistics. What conclusions might be drawn from the study using inferential statistics?
  Answer: Descriptive statistics involves statements such as "For unmarried men, approximately 70% were alive at age 65" and "for the married men, 90% were alive at age 65." A possible inference drawn from the study is that being married is associated with longer life for men

Variables

A **variable** is a characteristic or attribute measured on a sample or population elements and can assume different values.

There are two types of variables

- Quantitative Variable
- Qualitative Variable
- **Qualitative** – are variables that assume non-numeric values. Example: gender of a person,religious preference, geographic locations, place of birth, eye colour, political affiliation.

- **Quantitative** – are variables that assume numeric values. Example: heights, weights, and body temperatures.

- Quantitative variables can be further classified into two groups: **discrete** and **continuous**.

- **Discrete variables**: are variables that assume values that can be counted. Example: the number of children in a family, the number of students in a classroom.

- **Continuous variables**: are variables that assume all values between any two specific values. They areobtained by measuring. Example: temperature, between two points, weight of soil sample.

Scale of Measurement

- Data can also be classified by how they are categorized, counted, or measured. Example,area of residence (rural, suburban, or urban)?

- There are four scales of measuring data: and four common types of scales are used:
  - nominal scale
  - Ordinal scale
  - Interval scale
  - Ratio scale

- **Nominal Scale**: classifies data into mutually exclusive (nonoverlapping) categories and cannot be arranged in a certain order.

- Examples: political party (NDC, NPP, PNC, CPP. DFP, etc.), religion (Christianity, Judaism, Islam, etc.), and marital status (single, married, divorced, widowed, separated).

- The **Ordinal Scale**: classifies data into categories that can be ordered or ranked; however, precise differences between the ranks do not exist.

- Examples: severity of injury (fatal, serious, minor, no injury), academic performance(Excellent, very good, poor), socio-economic status (High, middle, low), Taste of food (Good, moderate, bad)

- **Interval Scale**: data can be ranked and precise differences between the ranks do exist; however, there is no meaningful zero.

- Example, Temperature: the difference between temperature of $100°c$ and $500°c$ indicates that one of warmer than the other. However, temperature of zero does not means that there is no temperature

- IQ of a person: the difference between IQ of 50 and 110 indicates that one is more intelligent than the other. However, IQ of zero does not means that the person has no intelligence.

- **Ratio level**: Possesses all the characteristics of interval scale, and there exists a true zero.

- Example: Distance between two locations, Weight of soil sample, Area of a landfill, Volume of a landfill

SAMPLING TECHNIQUES

Researchers use samples to collect data and information about a particular variable from a large population.

Four basic methods of sampling:

- Simple Random Sample
- Stratified Sample
- Cluster Sample
- Systematic Sample

- **Random Sample**: a sample in which all members of the population have an equal chance of being selected.

- Example: number each subject in the population. Then place numbered cards in a bowl, mix them thoroughly, and select as many cards as needed. The subjects whose numbers are selected constitute the sample.

- **Systematic Sample**: obtained by selecting every k_{th} member of the population where k is a counting number.

- For example, suppose there were 100 subjects in the population and a sample of 10 subjects were needed. Since $\frac{100}{10} = 10$, then k $= 10$, and every 10th subject would be selected; however, the first subject (numbered between 1 and 10) would be selected at random. Suppose subject 6 were the first subject selected; then the sample would consist of the subjects whose numbers were 6, 16, 26, etc., until 10 subjects were obtained.

- A **stratified sample**: obtained by dividing the population into subgroups or strata according to some characteristic relevant to the study. (There can be several subgroups.) Then subjects are selected from each subgroup.

- For example, suppose the vice-chancellor wants to learn how students feel about a certain issue. Furthermore, the vice-chancellor wishes to see if the opinions of first-year students differ from those of second-year, third-year, and fourth-year students. The vice-chancellor will randomly select students from each subgroup to use in the sample.

- A **cluster sample** is obtained by dividing the population into sections or clusters and then selecting one or more clusters and using all members in the cluster(s) as the members of the sample.

- For example, Suppose a researcher wishes to survey the average age of SHS students in Ashanti region. If there are 100 in the region, the researcher can randomly select 9 SHS from the 100 and interview all the Students of these SHS.

- Cluster sampling is used when the population is large or when it involves subjects residing in a large geographic area.

EXPERIMENTAL DESIGN

There are several different ways to classify statistical studies. Two types of studies will be considered: observational studies and experimental studies

- In an **observational study**, the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.

- In an **experimental study**, the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables

Data Collection

There are several different ways to classify statistical studies. Two types of studies will be considered: observational studies and experimental studies

- Data can be collected in a variety of ways. It is the most important part of statistical procedure because valid conclusions can only results from data which has been collected properly.

- If proper procedure is used to collect data then the issue of representativeness can be guaranteed. The use of wrong or faulty data collection methods would result in wrong conclusions because no good statistical tool can produce good results from wrongly collected data.

- There are **two** main types of data
  - **Primary data**
  - **Secondary data**

- **Primary Data**: This refers to data that was collected by the user. That is, data collected for the first time by the researcher for a defined purposes.

- Three of the most popular methods are: experiment, observation, surveys

- Advantages of using secondary data
  - It gives original research quality and does not carry bias or opinion of third parties
  - Ability to change the content or the course of study when ever needed
  - What needed is what is obtained (if it is well designed)

- Limitations of using secondary data.
  - Difficulty in designing suitable approach
  - Cost involve
  - Time consuming

- **Secondary Data**: This refers to data that was collected by someone other then the user. That is, data collected for other research purposes.

- Sources of secondary data includes: Government Records (census data, health records, educational institutes records), Private Organisations/Companies, Published materials (i.e. Books, Journals, websites)

- Advantages of using secondary data
  - It saves time and money,
  - It may be very accurate

- Limitations of using secondary data.
  - It may be incomplete
  - It may not be exactly what you need
  - It may not be consistent/reliable
  - It may be outdated

Choice of Data Source

- **Should I use secondary data or primary data?**

- Ask your self certain questions?
  - Will the data answer my research questions?
    - You must first decide what your research questions are
    - Then you need to decide what variables are needed to answer the questions scientifically

- If that data exist in secondary form, then use them to the extent you can, keeping in mind limitations

- But if it does not, and you are able to fund primary data collection, then it is the method of choice.

FREQUENCY DISTRIBUTIONS AND GRAPHS

Introduction

- When conducting a statistical study, the researcher must gather data for the particular variable under study.
- To describe situations, draw conclusions, or make inferences about events, the researcher must organize the data in some meaningful way. The most convenient method of organizing data is to construct a frequency distribution
- After organizing the data, the researcher must present them so they can be understood by those who will benefit from reading the study
- The most useful method of presenting the data is by constructing statistical charts and graphs. There are many different types of charts and graphs, and each one has a specific purpose

Descriptive Analysis of Data

- The analysis of data using descriptive methods involves:
  - Tabular presentation of data
  - Graphical presentation of
  - Numerical Summary

Raw Data

When data are collected in original form, they are called raw data. Example: the ages of the 50 wealthies people in the world.

$$49, 74, 54, 65, 48, 78, 52, 85, 60, 61, 57, 59, 56, 85, 81, 82, 56, 40, 71, 83,$$
$$38, 76, 69, 49, 68, 43, 81, 85, 57, 90, 73, 65, 68, 69, 37, 64, 77, 59, 61, 87,$$
$$81, 69, 78, 61, 43, 67, 79, 80, 69, 74.$$

Organizing Data

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

- Each raw data value is placed into a quantitative or qualitative category called a **class**. The **frequency** of a class then is the number of data values contained in a specific class.

Tabular Presentation of Data

A frequency distribution from the raw data is as presented below,

Class limits	Tally	Frequency
35–41	///	3
42–48	///	3
49–55	////	4
56–62	𝍸 𝍸	10
63–69	𝍸 𝍸	10
70–76	𝍸	5
77–83	𝍸 𝍸	10
84–90	𝍸	5
	Total	50

- It can be observed from the frequency distribution that the majority of the wealthy people in the study are 45 years old or older.
- The classes in this distribution are 35-41, 42-48, etc. These values are called class limits

Types of Frequency Distribution Table

Ungrouped table

Class	Frequency
1	3
2	5
3	12
4	4
5	8

Grouped table

Class	Frequency
10 – 14	13
15 – 19	15
20 – 24	12
25 – 29	14
30 – 34	18

Rules for constructing Frequency Distribution

To construct a frequency distribution, follow these rules:

- There should be between 5 and 20 classes
- The class must be mutually exclusive
- The class must be continuous
- The class must be exhaustive
- The class must be equal in width and size

Definitions

- Lower and upper class limit of a class represents the smallest and largest data value that can be included in the class. Lower and upper class limit for the class $1 - 10$ is 1 and 10 respectively

- Class boundaries– These numbers are used to separate the classes so that there are no gaps in the frequency distribution.

- Lower class boundary = average of upper limit of the previous class and the lower limit of the given class.
  Example: Lower boundary for the class $11 - 20 = \frac{10+11}{2} = 10.5$

- Upper class boundary = average of the upper limit of the class and the lower limit of the next class.
  Example: Lower boundary for the class $11 - 20 = \frac{20+21}{2} = 20.5$

- Class marks: They are the midpoints of the classes. They are obtained by averaging the limits
  Class mark of class $11 - 20 = \frac{1+10}{2} = 5.5$

- Class width = lower class limit of the next class - lower class limit of current class.
  Class Width = $11 - 1 = 10$

- Relative Frequency $= \frac{Frequency\ of\ the\ class}{Total\ Frequency}$. It describes the proportion of values falling in that class.

- Cumulative Frequency describes the number of observations that lies above (or below) a particular value in a data set.

Frequency Distribution

THE FREQUENCY TABLE

Class interval	frequency	Class mark	Class boundary	Cumulative frequency	Relative frequency	Cumulative relative frequency
1 – 10	2	5.5	0.5 – 10.5	2	0.10	0.10
11 – 20	3	15.5	10.5 – 20.5	5	0.15	0.25
21 – 30	1	25.5	20.5 – 30.5	6	0.05	0.30
31 – 40	3	35.5	30.5 – 40.5	9	0.15	0.45
41 – 50	2	45.5	40.5 – 50.5	11	0.10	0.55
51 – 60	4	55.5	50.5 – 60.5	15	0.20	0.75
61 – 70	5	65.5	60.5 – 70.5	20	0.25	1
	$\sum f = 20$				$\sum Rf = 1$	

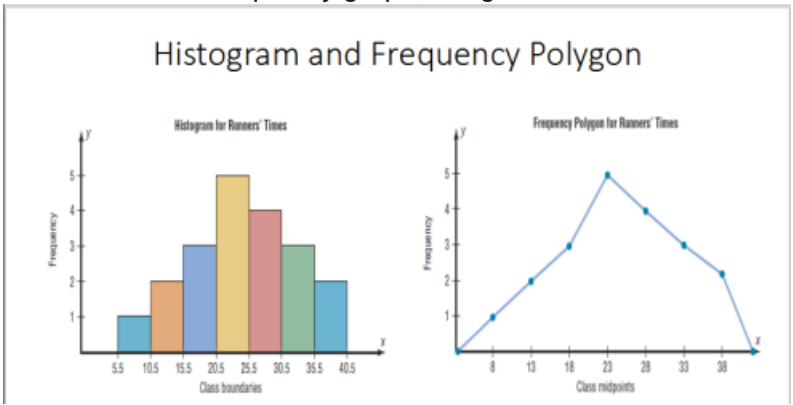
Constructing a Grouped Frequency Distribution

- **Step 1** Determine the classes.
  - Find the highest and lowest values
  - Find the range
  - Select the number of classes desired
  - Find the width by dividing the range by the number of classes and rounding up
  - Select a starting point (usually the lowest value or any convenient number less than the lowest value); add the width to get the lower limits
  - Find the upper class limits
  - Find the boundaries
- **Step 2** Tally the data
- **Step 3** Find the numerical frequencies from the tallies, and find the cumulative frequencies

Histograms, Frequency Polygons, and Ogives

The three most commonly used graphs in research are;

- The histogram
- The frequency polygon
- The cumulative frequency graph, or ogive



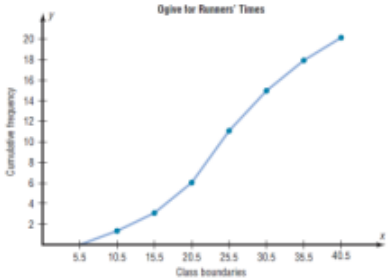
Drawing a Histogram

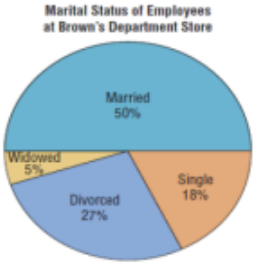
- **Step 1** Draw and label the x and y axes. The x axis is always the horizontal axis, and the y axis is always the vertical axis.
- **Step 2** Represent the frequency on the y axis and the class boundaries on the x axis
- **Step 3** Using the frequencies as the heights, draw vertical bars for each class.

Drawing a Frequency Polygon

- **Step 1** Find the midpoints of each class. Recall that midpoints are found by adding the upper and lower boundaries and dividing by 2
- **Step 2** Draw the x and y axes. Label the x axis with the midpoint of each class, and then use a suitable scale on the y axis for the frequencies.
- **Step 3** Using the midpoints for the x values and the frequencies as the y values, plot the points.
- **Step 4** Connect adjacent points with line segments. Draw a line back to the x-axis at the beginning and end of the graph.

Ogive and Pie Chart





Drawing an Ogive

- **Step 1** Find the cumulative frequency for each class.
- **Step 2** Draw the x and y axes. Label the x axis with the class boundaries. Use an appropriate scale for the y axis to represent the cumulative frequencies.
- **Step 3** Plot the cumulative frequency at each upper class boundary
- **Step 4** Starting with the first upper class boundary, connect adjacent points with line segments. Then extend the graph to the first lower class boundary, on the x axis.
- **Step 5** Cumulative frequency graphs are used to visually represent how many values are below a certain upper class boundary

Drawing a Pie Chart

The purpose of the pie graph is to show the relationship of the parts to the whole by visually comparing the sizes of the sections.

- **Step 1** Since there are 360° in a circle, the frequency for each class must be converted to a proportional part of the circle. This conversion is done by using the formula $\frac{Frequency\ for\ each\ class\ (f)}{sum\ of\ the\ frequencies\ (n)}$
- **Step 2** Each frequency is converted to a percentage using the formula $\% = \frac{f}{n} \times 100\%$
- **Step 3** Using a protractor and a compass, draw the graph, using the appropriate degree measures found in Step 1, and label each section with the name and percentages,

Stem and leave Plot

Stem-and-leaf Plot

Scores Earned by 50 Students in an Exam in Financial Accounting:

58	88	65	96	85
74	69	63	88	65
85	91	81	80	90
65	66	81	92	71
82	98	86	100	82
72	94	72	84	73
76	78	78	77	74
83	82	66	76	63
62	62	59	87	97
100	75	84	96	99

Stem	Leaf
5	8 9
6	2 2 3 3 5 5 5 6 6 9
7	1 2 2 3 4 4 5 6 6 7 8 8
8	0 1 1 2 2 2 3 4 4 5 5 6 7 8 8
9	0 1 2 4 6 6 7 8 9
10	0 0

Stem and Leave Plot

QUESTION

Example 1-9: A sample of the number of admissions to a psychiatric ward at a local hospital during the full phases of the moon is given below. Display the data using a stem-and-leaf plot with the leaves represented by the unit digits.

22	21	31	20	25	21	32	26	43	30	27
30	27	36	28	33	38	35	19	30	34	41

Solution: The stem-and-leaf display for the data is given in **Table 1-11.**

Table 1-11: Stem-and-Leaf Display for Example 1-9

STEM	LEAVES
1	9
2	0 1 1 2 5 6 7 7 8
3	0 0 0 1 2 3 4 5 6 8
4	1 3

MEASURES OF CENTRAL TENDENCY

- Measures of average are also called measures of central tendency and include the mean, median, mode.

- The measures that determine the spread of the data values are called measures of variation, or measures of dispersion. These measures include the range, variance, and standard deviation

- Finally, another set of measures is necessary to describe data. These measures are called measures of position. They tell where a specific data value falls within the data set or its relative position in comparison with other data values. The most common position measures are percentiles, deciles, and quartiles

- **Mean**(average) : sum all data entries and divide by the number of entries.
  Population Mean: $\mu = \frac{\sum x}{N}$
  Sample Mean : $\bar{x} = \frac{\sum x}{n}$
- **Median**: the middle value of an ordered set
- **Mode**:the data entry that occurs most frequent.

Example:
Find the mean, median and mode of the following data set

1. 200, 400, 300, 500, 400, 600, 700
2. 872, 397, 427, 388, 782, 397
3. 100, 101, 102, 103, 104, 105, 106
4. 250, 400, 350, 300, 300, 350, 450, 2000

Solution:

1. ordered data: 200,300, 400, 400, 500, 600,700
   mean = $\frac{200+300+400+400+500+600+700}{7}$ = $\frac{3100}{7}$ = 442.9
   median = 400
   mode = 400

2. ordered data: 388,397,397,427,782,872
   mean = $\frac{388+397+397+427+782+872}{6}$ = $\frac{3263}{6}$ = 442.9
   median = $\frac{397+427}{2} = \frac{824}{2} = 412$
   mode = 397

3. Mean = Median = 103. There is no mode in the data

4. 250, 300,300,350,350,400,450,2000 mean = $\frac{388+397+397+427+782+872}{6}$ = $\frac{3263}{6}$ = 442.9
   median = $\frac{350+350}{2} = 350$
   mode = 300 and 350. Thus the data is bimodal

So far which of the three measures of central tendency is

- greatly affected by outliers (extremely large or small values in a data set)?

- takes into account every entry of a data set? How would you decide which measure of central tendency would best represent a data set?

Mean of Grouped Data

Mean, $\bar{x} = \frac{\sum fx}{\sum f}$.

Class	Frequency, f	Class mark, x	fx
1 – 10	2	5.5	11
11 – 20	1	15.5	15.5
21 – 30	3	25.5	76.5
31 – 40	2	35.5	71
41 – 50	2	45.5	91
	$\sum f = 10$		$\sum fx = 265$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{265}{10} = 26.5$$

Weighted Mean

Mean, $\bar{x} = \frac{\sum wx}{\sum w}$; where w is the weight of each entry x .

Course code	Score, x	Credit hours, w	wx
OPT 153	85	3	255
CSM 183	70	3	210
ENGL 157	75	2	150
MATH 153	80	2	160
CHEM 159	65	3	195
		$\sum w = 13$	$\sum wx = 970$

$$\bar{x} = \frac{\sum wx}{\sum w} = \frac{970}{13} = 74.62$$

Properties of the mean

- The mean is found by using all the values of the data
- The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples
- The mean is used in computing other statistics, such as the variance
- The mean for the data set is unique and not necessarily one of the data values
- The mean cannot be computed for the data in a frequency distribution that has an open-ended class
- The mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations

Properties of the median

- The median is used to find the center or middle value of a data set
- The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution
- The median is used for an open-ended distribution
- The median is affected less than the mean by extremely high or extremely low values

Properties of the mean

- The mode is used when the most typical case is desired
- The mode is the easiest average to compute
- The mode can be used when the data are nominal or categorical, such as religious preference, gender, or political affiliation
- The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set

Distribution Shapes

- Frequency distributions can assume many shapes. The three most important shapes are positively skewed, symmetric, and negatively skewed. The figure shows histograms of each.

- When majority of the data values fall to the left of the mean, then data is said to be **positively skewed.** The “tail” is to the right.

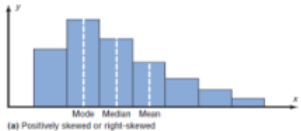
- When majority of the data values fall to the right of the mean, then data is said to be **negatively skewed**. The “tail” is to the left

- In a symmetric distribution, the data values are evenly distributed on both sides of the mean.

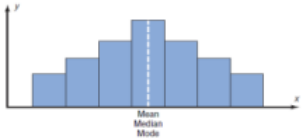
- In addition, when the distribution is unimodal, the mean, median, and mode are the same and are at the center of the distribution

Shapes(Skewness) of Frequency Distributions

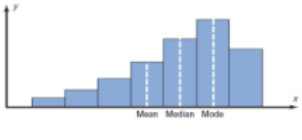
- **Symmetric:** mean = median = mode
- **Left skewed:** mean < median < mode
- **Right skewed:** mode < median < mean



(a) Positively skewed or right-skewed



(b) Symmetric



(c) Negatively skewed or left-skewed

MEASURES OF VARIATION

- **Range**: the difference between the maximum and the minimum data entries in the set.
- **Deviation** : the difference between the data entry, x and the mean of data set; $x - \mu$ or $x - \bar{x}$
- **Population Variance**: $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$
- **Sample variance** $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$
- **Population Standard deviation** $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$
- **Sample standard deviation** $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

MEASURES OF VARIATION

QUESTION

- For the data set: 10, 12, 13, 15, 25, 30

Range $= max - min = 30 - 10 = 20.$

$\bar{x} = \frac{10+12+13+15+25+30}{6} = 17.5$

x	Deviation, $x - \bar{x}$	$(x - \bar{x})^2$
10	$10 - 17.5 = -7.5$	$(-7.5)^2 = 56.25$
12	$12 - 17.5 = -5.5$	$(-5.5)^2 = 30.25$
13	$13 - 17.5 = -4.5$	$(-4.5)^2 = 20.25$
15	$15 - 17.5 = -2.5$	$(-2.5)^2 = 6.25$
25	$25 - 17.5 = 7.5$	$(7.5)^2 = 56.25$
30	$30 - 17.5 = 12.5$	$(12.5)^2 = 156.25$
	$\sum(x - \bar{x}) = 0$	$\sum(x - \bar{x})^2 = 325.5$

$s^2 = \frac{325.5}{6-1} = 65.1$

$s = \sqrt{65.1} = 8.07$

MEASURES OF VARIATION

QUESTION

- For the data set: 111, 112, 115, 117, 118, 119, 120

Range $= max - min = 120 - 111 = 9.$

$\bar{x} = \frac{111+112+115+117+118+119+120}{7} = 116\,.$

x	$(x-\bar{x})^2$
111	$(111-116)^2 = 25$
112	$(112-116)^2 = 16$
115	$(115-116)^2 = 1$
117	$(117-116)^2 = 1$
118	$(118-116)^2 = 4$
119	$(119-116)^2 = 9$
120	$(120-116)^2 = 16$
	$\sum(x-\bar{x})^2 = 72$

$$s^2 = \frac{72}{7-1} = 12$$

$$s = \sqrt{12} = 3.46$$

MEASURES OF VARIATION

- As previously stated, variances and standard deviations can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed. This information is useful in comparing two (or more) data sets to determine which is more (most) variable.
- The measures of variance and standard deviation are used to determine the consistency of a variable. For example, in the manufacture of fittings, such as nuts and bolts, the variation in the diameters must be small, or the parts will not fit together
- The variance and standard deviation are used to determine the number of data values that fall within a specified interval in a distribution. For example, Chebyshev's theorem (explained later) shows that, for any distribution, at least 75 % of the data values will fall within 2 standard deviations of the mean

Coefficient of Variation

- Whenever two samples have the same units of measure, the variance and standard for each can be compared directly. A statistic that allows you to compare standard deviations when the units are different, as in this example, is called the **coefficient of variation**
- The **coefficient of variation**, denoted by CVar, is the standard deviation divided by the mean. The result is expressed in percentage
- **For Samples**, CVar = $\frac{s}{\bar{x}} \times 100$
- **For Population**, CVar = $\frac{\sigma}{\mu} \times 100$

COEFFICIENT OF VARIATION

- The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

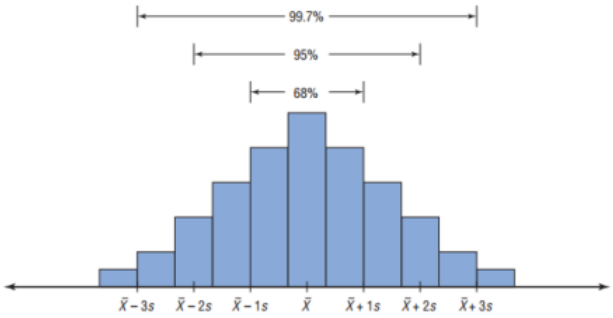
solution

The coefficient of variation are CVar = $\frac{s}{\bar{x}} = \frac{5}{87} \times 100$ CVar = $\frac{773}{5225} \times 100 = 14.8\%$ commissions Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

MEASURES OF VARIATION

- **The Empirical (Normal) Rule** – Chebyshev's theorem applies to any distribution regardless of its shape. However, when a distribution is bell-shaped (or what is called normal), the following statements, which make up the empirical rule, are true
- Approximately 68% of the data values will fall within 1 standard deviation of the mean
- Approximately 95% of the data values will fall within 2 standard deviations of the mean
- Approximately 99.7% of the data values will fall within 3 standard deviations of the mean

MEASURES OF VARIATION



MEASURES OF POSITION

- Quartiles – divide an ordered data set into approximately four equal parts. They are $Q_1, Q_2 \; and \; Q_3$.
- Deciles – divide an ordered data set into separately ten equal parts. They are $D_1, D_2, ...D_{99}$
- Percentiles – divide an ordered data set into approximately 100 equal parts. $P_1, P_2, ...P_{99}$

Compute $Q_1, Q_2 \; and \; Q_3$

1. 7, 18, 11, 6,59,17, 18, 54, 104 , 20, 31, 8, 10, 15, 19
2. 3, 6,15, 12, 8, 7
3. 25 , 33 ,4 , 37, 19, 15 , 20

Solution

1. Ordered data: 6, 7, 8, **10**, 11, 15, 17, **18**, 18, 19, 20, **31**, 54, 59, 104
   Q_2 = 18, Q_1 = 10 and Q_3 = 31

   IQR = Q_3 - Q_1 = 31 - 10 = 21

2. Ordered data: 3, 6, **7**, **8**, 12, 15
   Q_2 = $\frac{7+8}{2}$ = 7.5, Q_1 = 6 and Q_3 = 12

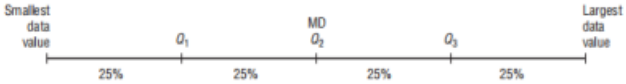
   IQR = Q_3 - Q_1 = 12 - 6 = 6

3. Ordered data: 4, 15, 19, **20**,25, 33, 37
   Q_2 = 20, Q_1 = 15 and Q_3 = 33

   IQR = Q_3 - Q_1 = 33 - 15 = 18

MEASURES OF POSITION

Quartiles and Deciles **Quartiles** divide the distribution into four groups seperated by $Q_1, Q_2\ and\ Q_3$. Note that Q_1 is the same as the 25th percentile; Q_2 is the 50th percentile or the median; Q_3 corresponds to the 75th percentile as shown:



MEASURES OF POSITION

- A data set should be checked for extremely high or extremely low values. These values are called **outliers**.
- An outlier is an extremely high or an extremely low data value when compared with the rest of the data values

PROCEDURE FOR CHECKING OUTLIERS

- Step 1 Arrange the data in order and find Q1 and Q3
- Step 2 Find the interquartile range: IQR Q3 - Q1
- Step 3 Multiply the IQR by 1.5
- Step 4 Subtract the value obtained in step 3 from Q1 and add the value to Q3
- Step 5 Check the data set for any data value that is smaller than Q1 - 1.5(IQR) or larger than Q3 - 1.5(IQR)

QUESTION

Check the the data set for outliers. 5, 6, 12, 13, 15, 18, 22, 50

BOX PLOT

Box Plot

Requires five-number summary namely:

- Minimum Entry
- First Quartile, Q_1
- Second quartile (Median), Q_2
- Third quartile, Q_3
- Maximum entry

BOX PLOT

Information obtained from a Box Plot

1. a. If the median is near the center of the box, the distribution is approximately symmetric
   b. If the median falls to the left of the center of the box, the distribution is positively skewed.
   c. If the median falls to the right of the center of the box, the distribution is negatively skewed

2. a. If the lines are about the same length,the distribution is approximately symmetric
   b. If the right line is larger than the left line, the distribution is positively skewed.
   c. If the left line is larger than the right line,the distribution is negatively skewed

BOX PLOT

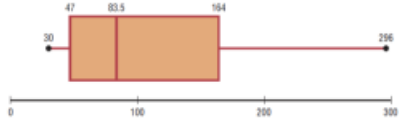
QUESTION

- The number of meteorites found in 10 states of the United States is 89, 47, 164, 296, 30, 215, 138, 78, 48, 39. Construct a boxplot for the data.

Answer

30, 39, 47, 48, **78**, **89**, 138, 164, 215, 296. $Min = 30$ and $Max = 296$;

$Q_2 = \frac{78+89}{2} = 83.5$, $Q_1 = 47$ and $Q_3 = 164$.



The distribution is somewhat positively skewed.

INTRODUCTION

- A cynical person once said, “The only two sure things are death and taxes.” This philosophy no doubt arose because so much in people’s lives is affected by chance

- From the time you awake until you go to bed, you make decisions regarding the possible events that are governed at least in part by chance.

- For example, should you carry an umbrella to work today? Will your car battery last until spring? Should you accept that new job?

PROBABILITY

Basic Concept of Probability

Probability as a general concept can be defined as the chance of an event occurring.

- **Probability experiment** – a chance process that leads to well–defined results called outcomes. Processes such as flipping a coin, rolling a die, or drawing a card from a deck are called probability experiments.
- **Outcome**– the result of a single trial of a experiment.
- **A trial means** flipping a coin once, rolling one die once, or the like. When a coin is tossed, there are two possible outcomes: head or tail. (Note:We exclude the possibility of a coin landing on its edge.) In the roll of a single die, there are six possible outcomes: 1, 2, 3, 4, 5, or 6. In any experiment, the set of all possible outcomes is called the **sample space**.

Basic Concept of Probability

- **Sample Space** – the set of all the possible outcomes of a probability experiment. The elements of the sample space are called sample points.

- **Event** - subset of the sample space. it is a collection of sample points with a common property, i.e consists of a set of outcomes of a probability experiment. An event with one outcome is called a **simple event**

ILLUSTRATIVE EXAMPLE

- **Probability Experiment** roll a die
- **Outcome**: 3
- **Sample Space** : $1, 2, 3, 4, 5, 6$
- **Event** :Die is event = $2, 4, 6$ Sample Spaces for various probability experiments
  - Toss a coin - H, T
  - Roll a die - $1, 2, 3, 4, 5, 6,$
  - Toss two coins – HH, HT, TH, TT
  - Toss a coin and roll a die – $H1, H2, H3, H4, H5, H6, T1, T2, T2, T4, T5, T6$

Basic Concept of Probability

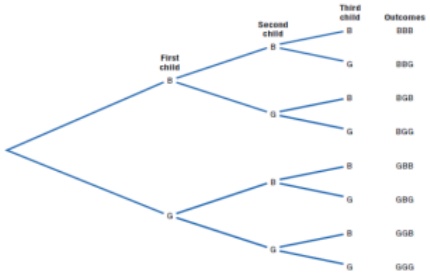
Question

- Find the sample space for rolling two dice
- Find the sample space for tossing a coin thrice

Another way to find all possible outcomes of a probability experiment is to use a **tree diagram**

TREED DIAGRAM

Find the sample space for the gender of the children if a family has three children. $BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG$



Types of Probability

There are three interpretations of Probability

- **Classical** Probability – outcomes in the sample space are equally likely occur.

$$P(E) = \frac{Number\ of\ outcomes\ in\ Event\ E}{Number\ of\ outcomes\ in\ sample\ space}$$

- **Empirical** Probability – relative frequency of an event.
$P(E) = \frac{Frequency\ of\ Event\ E}{Total\ frequency}$

- **Subjective** Probability – intuition, educated guesses and estimates. Eg. A doctor may feel a patient has a 90% chance of full recovery

PROBABILITY

QUESTION

You roll a six-sided die. Find the probability of each event:

1. Event A: rolling a 3
2. Event B: rolling a 7
3. Event C: rolling a number less than 5
4. Event D: rolling a prime number

Solution

Sample space: $S = \{1, 2, 3, 4, 5, 6\};\ n(S) = 6$

1. $A = \{3\};\ n(A) = 1;\ P(A) = \frac{n(A)}{n(S)} = \frac{1}{6}$.
2. $B = \{\};\ n(B) = 0;\ P(B) = \frac{n(B)}{n(S)} = \frac{0}{6} = 0$.
3. $C = \{1, 2, 3, 4\};\ n(C) = 4;\ P(C) = \frac{n(C)}{n(S)} = \frac{4}{6} = \frac{2}{3}$.
4. $D = \{2, 3, 5\};\ n(D) = 3;\ P(D) = \frac{n(D)}{n(S)} = \frac{3}{6} = \frac{1}{2}$.

QUESTION

IMPRESSION	NUMBER OF INDIVIDUALS
Positive	406
Negative	752
Neither	316
Don't know	30
	$\Sigma f = 1504$

- What is the probability that the next person surveyed has a positive overall impression?
- $P(positive) = \frac{f}{n} = \frac{406}{1504} = 0.27.$

Probability

Range of Probabilities

- The probability of an event E is between 0 and 1, inclusive.
  $0 \leq P(E) \leq 1$

Complementary Events

Complement of event **E** -the set of all outcomes in the sample space that are not included in **E** . it is denoted by $\bar{E}$

$P(\bar{E}) = 1 - P(E)$

Example: What is the probability that the next person surveyed does not have a positive overall impression?

- P(Positive) + $\frac{f}{n} = \frac{406}{1504} = 0.27$
- P(not Positive) = 1- 0.27 = 0.73

CONDITIONAL PROBABILITY AND MULTIPLICATION RULE

- **Conditional Probability** – the probability of event B occurring given that event A has already occurred;
- $P(B|A) = \frac{P(B \cap A}{P(A)}$

	Gene Present	Gene not present	Total
High IQ	33	19	52
Low IQ	39	11	50
TOTAL	72	30	102

- P(High IQ|Gene Present) = $\frac{33}{72} = 0.458$

CONDITIONAL PROBABILITY

A box contains black chips and white chips. A person selects two chips without replacement. If the probability of selecting a black chip *and* a white chip is $\frac{15}{56}$, and the probability of selecting a black chip on the first draw is $\frac{3}{8}$, find the probability of selecting the white chip on the second draw, *given* that the first chip selected was a black chip.

Solution

Let

B = selecting a black chip $\qquad$ W = selecting a white chip

Then

$$P(W|B) = \frac{P(B \text{ and } W)}{P(B)} = \frac{15/56}{3/8}$$

$$= \frac{15}{56} \div \frac{3}{8} = \frac{15}{56} \cdot \frac{8}{3} = \frac{\cancel{15}^{5}}{\cancel{56}_{7}} \cdot \frac{\cancel{8}^{1}}{\cancel{3}_{1}} = \frac{5}{7}$$

Hence, the probability of selecting a white chip on the second draw given that the first chip selected was black is $\frac{5}{7}$.

CONDITIONAL PROBABILITY

A recent survey asked 100 people if they thought women in the armed forces should be permitted to participate in combat. The results of the survey are shown.

Gender	Yes	No	Total
Male	32	18	50
Female	8	42	50
Total	40	60	100

Find these probabilities.

a. The respondent answered yes, given that the respondent was a female.
b. The respondent was a male, given that the respondent answered no.

Solution

Let

M = respondent was a male $\quad$ Y = respondent answered yes
F = respondent was a female $\quad$ N = respondent answered no

a. The problem is to find $P(Y|F)$. The rule states

$$P(Y|F) = \frac{P(F \text{ and } Y)}{P(F)}$$

CONDITIONAL PROBABILITY

The probability $P(F \text{ and } Y)$ is the number of females who responded yes, divided by the total number of respondents:

$$P(F \text{ and } Y) = \frac{8}{100}$$

CONDITIONAL PROBABILITY

The probability $P(F)$ is the probability of selecting a female:

$$P(F) = \frac{50}{100}$$

Then

$$P(Y|F) = \frac{P(F \text{ and } Y)}{P(F)} = \frac{8/100}{50/100}$$

$$= \frac{8}{100} \div \frac{50}{100} = \frac{\overset{4}{\cancel{8}}}{\underset{1}{\cancel{100}}} \cdot \frac{\overset{1}{\cancel{100}}}{\underset{25}{\cancel{50}}} = \frac{4}{25}$$

b. The problem is to find $P(M|N)$.

$$P(M|N) = \frac{P(N \text{ and } M)}{P(N)} = \frac{18/100}{60/100}$$

$$= \frac{18}{100} \div \frac{60}{100} = \frac{\overset{3}{\cancel{18}}}{\underset{1}{\cancel{100}}} \cdot \frac{\overset{1}{\cancel{100}}}{\underset{10}{\cancel{60}}} = \frac{3}{10}$$

Probability

- **Independent events** – two events are said to be independents if the occurrence of one event does not affect the probability of the occurrence of the other.
- **Two events are mutually exclusive events**if they cannot occur at the same time (i.e., they have no outcomes in common)
- $P(B \cap A) = P(B) or P(A \cap B) = P(A)$
- Events that are not independent are dependent

PROBABILITY

QUESTIONS

- Determine which events are mutually exclusive and which are not, when a single die is rolled
- Getting an odd number and getting an even number
- Getting a 3 and getting an odd number
- Getting an odd number and getting a number less than 4
- Getting a number greater than 4 and getting a number less than 4

PROBABILITY

The Multiplication Rule

- The probability that two events A and B will occur in sequence is
- $P(A\ and\ B) = (A \cap B) = P(A).P(B|A)$
- If A and B are independents events then
  $P(A\ and\ B) = P(A \cap B) = P(A).P(B)$

The Multiplication Rule

QUESTION

Example 1. A coin is tossed and a die is rolled. Find the probability of getting a head and then rolling a 6.

- The outcome on the coin does not affect the probability of rolling a 6 on the die. Thus, these two events are independent.
- $P(H \text{ and } 6) = P(H) \cdot P(6) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$.

Example 2. The probability that a particular knee surgery is successful is 0.85. If three knee surgeries are conducted, what is the probability that:

a) all three knee surgeries are successful?

b) none of the three surgeries is successful?

c) at least one of the surgeries is successful?

The Multiplication Rule

PROBABILITY

Solution

a) $P(3 \text{ surgeries are successful}) = 0.85 \times 0.85 \times 0.85 = 0.614.$

b) $P(success) = 0.85; \quad P(failure) = 1 - 0.85 = 0.15.$

Thus, $P(\text{none successful}) = 0.15 \times 0.15 \times 0.15 = 0.003.$

c) $P(\text{at least one successful}) = 1 - P(\text{none successful}).$

$= 1 - 0.003 = 0.997.$

The Multiplication Rule

QUESTION

Example 3. At a university in western Pennsylvania, there were 5 burglaries reported in 2003, 16 in 2004, and 32 in 2005. If a researcher wishes to select at random two burglaries to further investigate, find the probability that both will have occurred in 2004.

- In this case, the events are dependent since the researcher wishes to investigate two distinct cases.
- Hence the first case is selected and not replaced.

- $P(B_1 \text{ and } B_2) = P(B_1) \cdot P(B_2/B_1) = \frac{16}{53} \times \frac{15}{52} = \frac{60}{689} = 0.087.$

The Multiplication Rule

QUESTION

Example 4. World Wide Insurance Company found that 53% of the residents of a city had homeowner's insurance (H) with the company. Of these clients, 27% also had automobile insurance (A) with the company. If a resident is selected at random, find the probability that the resident has both homeowner's and automobile insurance with World Wide Insurance Company.

- $P(H \text{ and } A) = P(H) \cdot P(A/H) = 0.53 \times 0.27 = 0.1431.$

The Multiplication Rule

QUESTION

Try

A box contains 15 identical balls. Six of the balls are red, five blue and the rest white. If three balls are selected one after the other without replacement, find the probability that:

a) All three balls are red.

b) All three balls are not red.

c) First ball is red, second ball blue and the third ball white.

d) There is at least one white ball.

The Multiplication Rule

QUESTION

Try

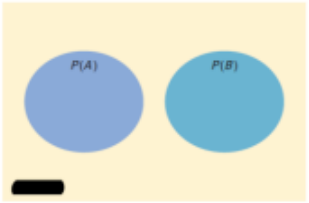
An urn contains 3 red balls, 2 blue balls, and 5 white balls. A ball is selected and its colour noted, then it is replaced. A second ball is selected and its colour noted. Find the probability of each of these.

- Selecting 2 blue balls
- Selecting 1 blue ball and then 1 white ball
- Selecting 1 red ball and then 1 blue ball

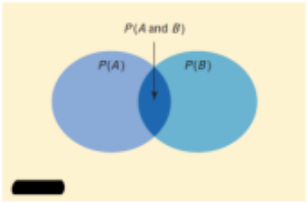
The Multiplication Rule

Mutually Exclusive Events and the Addition Rule

- **Mutually Exclusive –** two events A and B cannot occur at the same time.
- $P(A \cap B) = 0$.



(a) Mutually exclusive events



(b) Nonmutually exclusive events

ADDITION RULE

The Addition Rule

The probability that two events A or B will occuris

- $P(A\ or\ B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If A and B are mutually exclusive then
  $P(A\ or\ B) = P(A \cup B) = P(A) + P(B)$
  Example 1 You roll a die. Find the probability of obtaining a number less than 3 or an odd number.
  $S = \{1, 2, 3, 4, 5, 6\}; A = \{1, 2\}; B = \{1, 3, 5\}; A \cap B = \{1\}$
  - $P(A\ or\ B) = P(A\ \cup\ B) = P(A) + P(B)\ - P(A \cap B)$
  - $P(A \cup B) = \frac{2}{6} + \frac{3}{6} - \frac{1}{3} = \frac{2}{3}$

ADDITION RULE

Question

- A day of the week is selected at random. Find the probability that it is a weekend day.

ADDITION RULE

QUESTION

$Example\ 2$: A blood bank catalogues the types of blood given by donors during the last five days.

	O	A	B	AB	TOTAL
Rh^+	156	139	37	12	344
Rh^-	28	25	8	4	65
TOTAL	184	164	45	16	409

A donor is selected at random. Find the probability that the donor has
a) type **O** or type **A** blood
b) type **B** or **Rh negative**

ADDITION RULE

Solution

a) $P(O \text{ or } A) = P(O) + P(A) = \frac{184}{409} + \frac{164}{409}$

b) $P(B \text{ or } Rh^-) = P(B) + P(Rh^-) - P(B \text{ and} Rh^-)$
= $\frac{45}{409} + \frac{65}{409} - \frac{8}{409} = \frac{102}{409}$

$Example\ 3$: At a political rally, the are 20 Republicans, 13 Democrats and 6 Independents. If a person is selected at random, find the probability that he or she is either a Democratic or an Independent.

Ans: $n(S) = 20 + 13 + 6 = 39$

$P(D \text{ or } I) = P(D) + P(I)$

$\frac{13}{39} + \frac{6}{39} = \frac{19}{39}$

ADDITION RULE

QUESTIONS

$Example\ 4$: In the hospital unit there are 8 nurses and 5 physicians; 7 nurses and 3 physicians are females. If a staff person is selected, find the probability that the subject is a nurse or a male.

Ans:

Staff	Females	Males	Total
Nurses	7	1	8
Physicians	3	2	5
Total	10	3	13

$P(N\ or\ M) = P(N) + P(M) - P(N\ and\ M)$

= $\frac{8}{13} + \frac{3}{13} - \frac{1}{13} = \frac{10}{13}$

ADDITION RULE

Test 1

1. If A and B are independent events with P(A) = 0.3 P(B) = 0.6, find $P(A \cup B$
2. If P(A) = 0.6, P(B) = 0.3 and P(A|B) = 0.4. find $P(A \cup B)$
3. If P(A) = 0.6, P(B) = 0.3 and P(A|B) = 0.4, find $P(A^c)$
4. In a three-child family, what is the probability that there are at least two girls?
5. If two events are mutually exclusive, then they are independent.
   TRUE or FALSE

PROBABILITY

- **Fundamental Counting Principle**:If one event can occur in **m** ways and a second event can occur in**n** ways, the number of ways the two events can occur in sequence **m** $\times$ **n**
- Can be extended for any number of events occurring in sequence.
  Example: You are purchasing a new car. The manufacturers, car sizes and colours are listed
  Manufacturer: Ford, GM, Honda
  Car Size: compact, midsize
  Colour: White(W), Red(R), Black(B), Green(G)

  How many different ways can you select one manufacture, one car and one colour?

PROBABILITY

Probability And Counting

Solution

There are three choices of manufacturers, two car sizes and four colours. Using the fundamental counting principle:

$$3 \times 2 \times 4 = 24 \text{ ways}.$$

Factorial Notation

For any counting number n

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1.$$

Example:

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120.$$

$$4! = 4 \times 3 \times 2 \times 1 = 24.$$

By definition; $0! = 1$.

PROBABILITY

COUNTING TECHNIQUES

Permutations

- A **permutation** is an arrangement of $\boldsymbol{n}$ objects in a specific order.
- The arrangement of $\boldsymbol{n}$ objects in a specific order using $\boldsymbol{r}$ objects at a time
- $_nP_r = \frac{n!}{(n-r)!}$; where $r \leq n$.

Example 1: In how many ways can the letters of the word $\boldsymbol{CAR}$ be arranged in order?

Ans: $3! = 3 \times 2 \times 1 = 6$. $S = \{\boldsymbol{CAR}, \boldsymbol{CRA}, \boldsymbol{ACR}, \boldsymbol{ARC}, \boldsymbol{RCA}, \boldsymbol{RAC}\}$

Example 2: In how many ways can seven athletes finish first, second and third in an Olympic?

Ans: $_7P_3 = \frac{7!}{(7-3)!} = 7 \times 6 \times 5 = 210.$

PROBABILITY

QUESTION

$Example\ 3$: Find the number of ways of forming four-digit codes in which no digit is repeated.

Ans: n = 10 and r = 4

$^{10}P_4 = \frac{10!}{(10-4)!} = 10 \times 9 \times 8 \times 7 = 5040 ways$

$Example\ 4$: Suppose a business owner has a choice of 5 locations in which to establish her business. She decides to rank each location according to certain criteria, such as price of the store and parking facilities. How many different ways can she rank the 5 locations?

Ans: n = 5 and r = 5

$^{5}P_5 = \frac{5!}{(5-5)!} = 5! = 120 ways$

PROBABILITY

COUNTING TECHNIQUES

Distinguishable Permutations

- The number of distinguishable permutations of n objects where n_1 are of one type, n_2 are of another type, and so on is given by:
- $\frac{n!}{n_1! \times n_2! \times n_3! \cdots n_k!}$; where $n_1 + n_2 + n_3 + \cdots + n_k = n$.

Example 5: In how many ways can the letters of the word ***MISSISSIPPI*** be arranged?

Ans: $\frac{11!}{4! \times 4! \times 2!} = 34{,}650$ ways.

Example 6: In how many ways can the letters of the word ***STATISTICS*** be arranged?

Ans: $\frac{10!}{3! \times 3! \times 2!} = 50{,}400$ ways.

PROBABILITY

COUNTING TECHNIQUES

Combinations

- A selection of $\boldsymbol{r}$ objects from a group of $\boldsymbol{n}$ objects with no regard to order.
- $${}_nC_r = \binom{n}{r} = \frac{n!}{(n-r)! \times r!} \text{ ; where } r \leq n.$$

Example 6: A newspaper editor has received 8 books to review. He decides that he can use 3 reviews in his newspaper. How many different ways can these 3 reviews be selected?

Ans: ${}_8C_3 = 56.$

Example 7: In a club there are 7 women and 5 men. A committee of 3 women and 2 men is to be chosen. How many different possibilities are there?

Ans: ${}_7C_3 \cdot {}_5C_2 = 35 \times 10 = 350.$

PROBABILITY

Finding Probabilities

$Example\ 1$: Your exam ID number consists of 7 digits. Each number can be from 0 to 9 and each digit can be repeated. What is the probability of getting your ID number when randomly generating seven digits?

Solution

- Each digit can be repeated
- There are 10 choices for each of the 7 digits
- Using the fundamental counting principle, there are $10 * 10 * 10 * 10 * 10 * 10 * 10 = 10^7 = 10,000,000$ possible identification numbers
- Only one of those numbers corresponds to your ID number $P(Your\ ID\ number) = \frac{1}{10,000,000}$

PROBABILITY

QUESTION

Example 2: You have 11 letters consisting of one M, four I's, four S's and two P's. If the letters are randomly arranged in order, what is the probability that the arrangement spells the word Mississippi?

Solution
There is only one favourable outcome.

There are $\frac{11!}{1!\times 4!\times 4!\times 2!} = 34,650$ distinguishable permutations of the given letters

$$P(MISSISSIPPI) = \frac{1}{34650} = 0.000029$$

PROBABILITY

QUESTIONS

Example 3: A student advisory board consists of 17 members. Three members serve as the board's chair, secretary and webmaster. Each member is equally likely to serve any of the positions. What is the probability of selecting at random the three members that hold each position?

Solution

There is only one favourable outcome.

There are $_{17}P_3 = 17 \times 16 \times 15 = 4080$

ways the three positions can be occupied.

$\therefore P(selecting\ the\ 3\ members) = \frac{1}{4080} = 0.000245.$

PROBABILITY

QUESTIONS

Example 4: A board consists of 12 men and 8 women. If a committee of 3 members is to be formed what is the probability that

a) it includes at least one woman?
b) it includes more women than men?

Solution

$n = 12M + 8W = 20;\ \ r = 3$

The number of ways of forming the committee of 3 from 20

$$\binom{20}{3} = 1140$$

PROBABILITY

SOLUTIONS

a) Probability of at least one woman

$$= P(1W \text{ and } 2M) + P(2W \text{ and } 1M) + P(3W)$$

$$\frac{\binom{8}{1}\cdot\binom{12}{2}+\binom{8}{2}\cdot\binom{12}{1}+\binom{8}{3}}{1140} = 0.8070.$$

b) Probability of more women than men

$$= P(2W \text{ and } 1M) + P(3W)$$

$$\frac{\binom{8}{2}\cdot\binom{12}{1}+\binom{8}{3}}{1140} = 0.3439.$$

PROBABILITY

QUESTIONS

$Example\ 5$: A box contains 6 red, 3 white and 5 blue balls. If three balls are drawn at random, one after the other without replacement, find the probability that:

a) all are red
b) two are red and one is white
c) at least one is red
d) one of each colour

Solution

$$n = 6R + 3W + 5B = 14; r = 3$$

PROBABILITY

SOLUTIONS

a) $P(all\ 3\ R) =$

$$P(1^{st}R) \times P\left(2^{nd}R/1^{st}R\right) \times P\left(3^{rd}R/(1^{st}and\ 2^{nd}R)\right)$$

$$= \frac{6}{14} \times \frac{5}{13} \times \frac{4}{12} = \frac{\binom{6}{3}}{\binom{14}{3}} = 0.0549\,.$$

b) $P(2R \cap 1W) = \dfrac{\binom{6}{2}\cdot\binom{3}{1}}{\binom{14}{3}} = 0.1236\,.$

PROBABILITY

SOLUTIONS

c) $P(at\ least\ 1R) = 1 - P(no\ red) = 1 - \frac{\binom{8}{3}}{\binom{14}{3}} = 0.8462$.

d) $P(1R \cap 1W \cap 1B\ ) = \frac{\binom{6}{1}\cdot\binom{3}{1}\cdot\binom{5}{1}}{\binom{14}{3}} = 0.2473$.

PROBABILITY

QUESTIONS

$Example\ 6$: A food manufacturer is analyzing a sample of 100 cashew nuts for the presence of a toxin. If four nuts are randomly selected from the sample, what is the probability that exactly one nut contains a dangerously high level of the toxin?

Solution

$$P(1\ toxic\ of\ nut) = \frac{^{3}C_{1} * ^{97}C_{3}}{^{100}C_{4}} = 0.1128$$

PROBABILITY

TRY THESE

- In how many ways can the letters of the **MATHS** be arranged.
- In how many ways can the letters of **MEASURING** be arranged.
- In how many ways can **5 different objects** be arranged taking two at at time.
- In the managing of a committee of a society , there are nine members. In how many ways can a **chairman**, a **vice chairman** and a **treasurer** be selected from amongst them if a person is eligible for one post only

PROBABILITY

TRY THESE

- In how many ways can the letters of **CALCULUS** be arranged.
- In how many ways can the letters of **STATISTICS** be arranged
- Find the number of ways can the in which the letters of the word **ADDING** can be arranged if
  1. the two D's are together.
  2. the two D's are separated

PROBABILITY

- Find the number of ways in which the letters of the word **DEFLATED** can be arranged if

    1. the two E's are together
    2. the two E's are separated

- In how many ways can five children be seated around a circular table.
- In how many ways can six different coloured beads be placed on a ring.
- In how many ways can five beads, chosen from eight different beads be threaded on to a ring

PROBABILITY

TRY THESE

- In how many ways can a committee can a committee of 6 men and 4 women to be formed from a grouped of 8 men and 7 women.

- There are 15 boys and 10 girls, out of whom a committee of 3 boys and two girls is to be formed. Find the number of ways this can be done if

    1. there is no restriction
    2. a particular boy is included
    3. a particular girl is excluded

PROBABILITY

A committee of 6 members is to be formed from a teaching staff 10 women and 4 men

1. Find the number of ways of forming the committee
2. What is the probability that the committee
    - consists of only women
    - includes exactly three men
    - includes at least one man

A bag 8 red, 3 white and 9 black balls. If three balls are drawn at random determine the probability that

1. all 3 are red
2. all 3 are white

RANDOM VARIABLES

DISCRETE PROBABILITY DISTRIBUTIONS

- **Random Variable**:a variable whose values are determined by chance. They are represented by block letters like X or Y.
- **Discrete random variable** has a countable number of possible values.
  Example 1: Let X be the number of chairs in a classroom.
  Example 2: Let Y be number of phone calls made in a day.
- **Continuous random variable** can assume all values in the interval between any two given values. It is obtained from data that can be measured rather than counted.
  Example 3: Let X be the temperature within 24 hours. Example 4: Let Y be the height of MATH 153 students

RANDOM VARIABLES

PROBABILITY DISTRIBUTION

- A **discrete probability distribution** consists of the values a discrete random variable can assume and the corresponding probabilities of the values.
- Below is the probability distribution for rolling a single die:

Outcome X	1	2	3	4	5	6
Probability P(X)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

RANDOM VARIABLES

QUESTION

- The following table is the frequency distribution for the test scores of 100 students in a Statistics class. Use it to construct a probability distribution

X	f
0	20
1	25
2	15
3	10
4	20
5	10
	$\sum f = 100$

RANDOM VARIABLES

- Solution

X	f	$P(X) = \frac{f}{\sum f}$
0	20	0.2
1	25	0.25
2	15	0.15
3	10	0.10
4	20	0.20
5	10	0.10
	$\sum f = 100$	$\sum P(X) = 1$

- The probability distribution is given below:

X	0	1	2	3	4	5
P(X)	0.20	0.25	0.15	0.10	0.20	0.10

RANDOM VARIABLES

- Construct a probability distribution for the number of times a head shows up when a coin is tossed twice.
  Let X represent the number of heads that show up.

$S = \{HH, HT, TH, TT\}$
$X = \{0, 1, 2\}$
$P(X = 0) = P(TT) = \frac{1}{4} = 0.25$
$P(X = 1) = P(HT \; or \; TH) = \frac{2}{4} = 0.5$
$P(X = 2) = P(HH) = \frac{1}{4} = 0.25$

X	0	1	2
P(X)	0.25	0.50	0.25

RANDOM VARIABLES

- Construct a probability distribution for the number of girls of a family with three children.
  Let X represent the number of girls.
  $S = \{BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG\}$
  $X = \{0, 1, 2, 3\}$
  $P(X = 0) = \frac{1}{8} = 0.125$
  $P(X = 1) = \frac{3}{8} = 0.375$
  $P(X = 2) = \frac{3}{8} = 0.375$
  $P(X = 3) = \frac{1}{8} = 0.125$

X	0	1	2	3
P(X)	0.125	0.375	0.375	0.125

RANDOM VARIABLES

Two **requirements** for a probability distribution:

1. $0 \leq P(X) \leq 1$
2. $\sum P(X) = 1$

Example

Use the probability distribution given to answer the questions that follow:

X	0	1	2	3	4	5
P(X)	0.15	0.35	α	0.10	0.05	0.20

i. Find the value of α

ii. Find $P(X \leq 2)$

RANDOM VARIABLES

Solution

i. $\sum P(X) = 1$;

0.15+0.35+ α +0.10+0.05+0.20=1

α=1-0.85=0.15

ii. $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

0.15+0.35+0.15=0.65

RANDOM VARIABLES

QUESTIONS

Determine whether each distribution is a probability distribution

a.

X	1	2	3	4
P(X)	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{9}{16}$

b.

X	0	2	4	6
P(X)	-1.0	1.5	0.3	0.2

c.

X	0	5	10	15	20
P(X)	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

d.

X	2	3	7
P (X)	0.5	0.3	0.4

RANDOM VARIABLES

Mean and Variance of Discrete Random Variables

- Mean $\mu = \sum XP(X)$
- Variance $\sigma^2 = \sum (X-\mu)^2 P(X)$

- Example: Compute the mean,variance and standard deviation:

X	0	1
P (X)	0.5	0.5

Solution

$$\mu = 0(0.5) + 1(0.5) = 0.5$$
$$\sigma^2 = 0.5(0-0.5)^2 + 0.5(1-0.5)^2 = 0.25$$
$$\sigma = \sqrt{0.25} = 0.5$$

RANDOM VARIABLES

- Find the mean,variance and standard deviation of the number of girls in a family with two children.

Solution

$S = \{BB, BG, GB, GG\}$

$P(X = 0) = \frac{1}{4} = 0.25; P(X = 1) = \frac{2}{4} = 0.5; P(X = 2) = \frac{1}{4} = 0.25$

X	0	1	2
P(X)	0.25	0.5	0.25

$\mu = 0(0.25) + 1(0.5) + 2(0.25) = 1$

$\sigma^2 = 0.25(0.1)^2 + 0.5(1-1)^2 + 0.25(2-1)^2 = 0.5$

$\sigma = \sqrt{0.5} = 0.7071$

RANDOM VARIABLES

EXPECTATION

- The **expected value** of a discrete random variable of a probability distribution is the theoretical average of the variable.
- $E(X) = \mu = \sum XP(X)$

Examples

- One thousand tickets are sold at $1 each for a laptop valued at $350. What is the expected value of the gain if you purchase one ticket?

	Win	Lose
Gain X	$349	-$1
Probability of Gain P(X)	$\frac{1}{1000}$	$\frac{999}{1000}$

$E(X) = \$349(\frac{1}{1000}) + (-\$1)(\frac{999}{1000}) = -\$0.65$

RANDOM VARIABLES

QUESTION

2. One thousand tickets are sold at $1 each for four prizes of $100, $50, $25 and $10. After each prize drawing, the winning ticket is then returned to the pool of tickets. What is the expected value if you purchase two tickets?

Solution

Gain X	$98	$48	$23	$8	-$2
Probability P(X)	$\frac{2}{1000}$	$\frac{2}{1000}$	$\frac{2}{1000}$	$\frac{2}{1000}$	$\frac{992}{1000}$

$E(X) = \$900(\frac{2}{1000}) + \$48(\frac{2}{1000}) + \$23(\frac{2}{1000}) + \$8(\frac{2}{1000}) - \$2(\frac{992}{1000})$

= $-\$1.63$

RANDOM VARIABLES

Test 2

1. Use the probability distribution given below to answer the questions.

X	0	2	5
P(X)	0.2	0.3	2y

2. If 1,000 raffle tickets were sold for a phone worth 400 cedis, what is the expected value of the raffle? Would it be wise to spend more than 40 pesewas on the ticket?

RANDOM VARIABLES

QUESTION 1

A box contains 5 balls. Two are numbered 3, one is numbered 4, and two are numbered 5. The balls are mixed and one is selected at random. After a ball is selected, its number is recorded. Then it is replaced. If the experiment is repeated many times, find the variance and standard deviation of the numbers on the balls

RANDOM VARIABLES

QUESTION 2

A talk radio station has four telephone lines. If the host is unable to talk (i.e., during a commercial) or is talking to a person, the other callers are placed on hold.
When all lines are in use, others who are trying to call in get a busy signal. The probability that 0, 1, 2, 3, or 4 people will get through are 0.18, 0.34, 0.23, 0.21, 0.04 respectively.
Find the variance and standard deviation for the distribution

Thank You

Doubts and Suggestions

saasgyam@gmail.com or asante.gyamerah@knust.edu.com

STATISTICAL METHODS 1

MATH 153

GYAMERAH, Samuel Asante (Ph.D.)[1]

[1]Department of Statistics and Actuarial Science
Kwame Nkrumah University of Science and Technology

2021/2022

Department of Statistics and Actuarial Science, KNUST