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

2020/2021

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 2182 students who accepted admission offers to KNUST in 2009 an average cut-of points of 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 the United States found that 54% drink coffee daily. 54% is a parameter. TRUE or FALSE?

Branches of Statistics

- **Descriptive Statistics** – involves organizing, summarizing and displaying data

- **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

  Here, the statistician tries to make inferences from samples to populations. Inferential statistics uses probability, i.e., the chance of an event occurring

Branches of Statistics

- Decide which part of the study represents the descriptive branch of statistics. What conclusions might be drawn from the study using inferential statistics?

- 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. 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

Data Classification

Types of Variables

- **Qualitative** – variables that can be placed into distinct categories, according to some characteristic or attribute. For example, if subjects are classified according to gender (male or female), then the variable gender is qualitative. Other examples of qualitative variables are religious preference, geographic locations, place of birth, eye colour, political affiliation.

- **Quantitative** – are numerical and can be ordered or ranked. For example, the variable age is numerical, and people can be ranked in order according to the value of their ages. Other examples of quantitative variables are heights, weights, and body temperatures.

INTRODUCTION TO STATISTICS

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

- **Discrete variables** can be assigned values such as $0, 1, 2, 3$ and are said to be countable. Examples of discrete variables are the number of children in a family, the number of students in a classroom, and the number of calls received by a switchboard operator each day for a month.

- **Continuous variables**, by comparison, can assume an infinite number of values in an interval between any two specific values. Temperature, for example, is a continuous variable, since the variable can assume an infinite number of values between any two given temperature. They often include fractions and decimals

INTRODUCTION TO STATISTICS

- In addition to being classified as qualitative or quantitative, variables can be classified by how they are categorized, counted, or measured. For example, can the data be organized into specific categories, such as area of residence (rural, suburban, or urban)?

- Can the data values be ranked, such as first place, second place, etc.? Or are the values obtained from measurement, such as heights, IQs, or temperature? This type of classification—i.e., how variables are categorized, counted, or measured—uses **measurement scales**, and four common types of scales are used: **nominal, ordinal, interval, and ratio**

INTRODUCTION TO STATISTICS

- **The nominal level** of measurement classifies data into mutually exclusive (nonoverlapping) categories in which no order or ranking can be imposed on the data.

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

- The **ordinal level** of measurement classifies data into categories that can be ordered or ranked; however, precise differences between the ranks do not exist.

- Examples: from student evaluations, guest speakers might be ranked as superior, average, or poor. Floats in a homecoming parade might be ranked as first place, second place, etc

INTRODUCTION TO STATISTICS

- **The interval level** of measurement ranks data, and precise differences between units of measure do exist; however, there is no meaningful zero.

- For example, many standardized psychological tests yield values measured on an interval scale. IQ is an example of such a variable. There is a meaningful difference of 1 point between an IQ of 109 and an IQ of 110. Temperature is another example of interval measurement, since there is a meaningful difference of 1 between each unit, such as $72°C$ and $73°C$

- **Note** One property is lacking in the interval scale: There is no true zero. For example, IQ tests do not measure people who have no intelligence. For temperature, 0F does not mean no heat at all

INTRODUCTION TO STATISTICS

- **The ratio level** of measurement possesses all the characteristics of interval measurement, and there exists a true zero. In addition, true ratios exist when the same variable is measured on two different members of the population.

- Examples of ratio scales are those used to measure height, weight, area, and number of phone calls received. Ratio scales have differences between units (1inch,1pound, etc.) and a true zero. In addition, the ratio scale contains a true ratio between values. For example, if one person can lift 200 pounds and another can lift 100 pounds, then the ratio between them is 2 to 1. Put another way, the first person can lift twice as much as the second person

LEVELS OF MEASUREMENT

Sampling Techniques Researchers use samples to collect data and information about a particular variable from a large population. To obtain samples that are unbiased (that give each subject in the population an equally likely chance of being selected), Statisticians use four basic methods of sampling:

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

SAMPLING TECHNIQUES

- A **random sample** is a sample in which all members of the population have an equal chance of being selected.

- Random samples are selected by using chance methods or random numbers. One such method is to 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.

SAMPLING TECHNIQUES

- A **systematic sample** is a sample obtained by selecting every k_{th} member of the population where k is a counting number.

- For example, suppose there were 2000 subjects in the population and a sample of 50 subjects were needed. Since $\frac{2000}{50} = 40$, then k $= 40$, and every 40th subject would be selected; however, the first subject (numbered between 1 and 40) would be selected at random. Suppose subject 12 were the first subject selected; then the sample would consist of the subjects whose numbers were 12, 52, 92, etc., until 50 subjects were obtained.

SAMPLING TECHNIQUES

- A **stratified sample** is a 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.

SAMPLING TECHNIQUES

- 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 apartment dwellers in a large city. If there are 10 apartment buildings in the city, the researcher can select at random 2 buildings from the 10 and interview all the residents of these buildings. Cluster sampling is used when the population is large or when it involves subjects residing in a large geographic area. For example, if one wanted to do a study involving the patients in the hospitals in Ashanti Region, it would be very costly and time-consuming to try to obtain a random sample of patients since they would be spread over a large area. Instead, a few hospitals could be selected at random, and the patients in these hospitals would be interviewed in a cluster.

EXPERIMENTAL DESIGN

There are several different ways to classify statistical studies. Four 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

- Other data collection techniques are **simulation** and **surveys**

DESCRIPTIVE ANALYSIS OF DATA

- 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

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

- The three most commonly used graphs in research are;
  - The histogram
  - The frequency polygon
  - The cumulative frequency graph, or ogive

DATA COLLECTION

Frequency Distribution and Graphs

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

Suppose a researcher wished to do a study on the ages of the top 50 wealthiest people in the world.

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

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

DATA COLLECTION

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
	$\sum f = 20$				$\sum Rf = 1$	

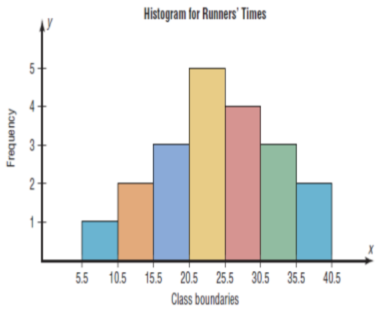
DATA COLLECTION

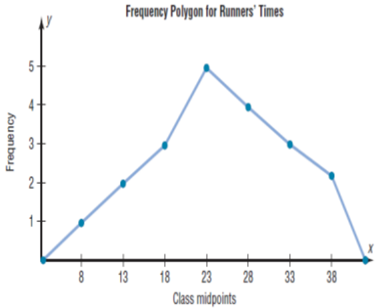
- class width $= 11 - 1 = 10$
- class mark $= \frac{1+10}{2} = 5.5$
- lower class limit for the class $1 - 10$ is 1
- upper class limit for the class $1 - 10$ is 10
- lower class boundary for the class $1 - 10$ is 0.5
- upper class boundary for the class $1 - 10$ is 10.5
- class boundary $= \frac{11-10}{2} = 0.5$;

  subtract and add 0.5 to the lower and upper class limits respectively

DESCRIPTIVE ANALYSIS OF DATA

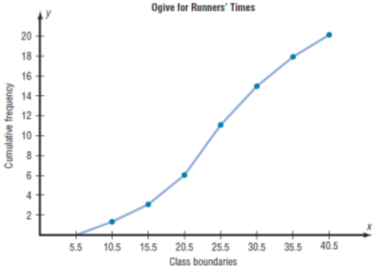
Histogram and Frequency Polygon

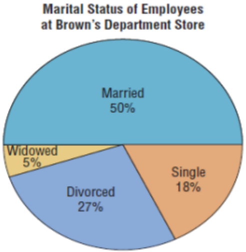




DESCRIPTIVE ANALYSIS OF DATA

Ogive and Pie Chart





DESCRIPTIVE ANALYSIS OF DATA

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

DESCRIPTIVE ANALYSIS OF DATA

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

MEASURES OF CENTRAL TENDENCY

- **Mean** (average)– sum all data entries and divide by the number of entries.

  Population Mean: $\mu = \frac{\Sigma x}{N}$ Sample mean: $\bar{x} = \frac{\Sigma x}{n}$
- **Median**– the middle value of an ordered data set.
- **Mode**–the data entry that occurs most frequent.

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. $205 , 300 , 300 , 350 , 350 , 400 , 450 , 2000$

solution

1. order 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. order data: $388\,,397\,,397\,,427\,,782\,,872$

   mean$= \frac{388+397+397+427+782+872}{6} = \frac{3263}{6} = 543.8$

   median $= \frac{397+427}{2} = \frac{824}{2} = 412$

   mode$= 397$

CONT"

1. Mean $=$ Median $= 103$. There is **no mode** in the data.
2. $250\,,300\,,300\,,350\,,350\,,400\,,450\,,2000$

   Mean $= \frac{205+300+300+350+350+400+450+2000}{8} = \frac{4400}{8} = 550$

   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

- **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?

MEASURES OF CENTRAL TENDENCY

Mean of Grouped Data

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

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
	$\sum f = 10$		$\sum fx = 265$

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

MEASURES OF CENTRAL TENDENCY

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 MODE

- 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

MEASURES OF CENTRAL TENDENCY

- 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.

- In a positively skewed or right-skewed distribution, the majority of the data values all to the left of the mean and cluster at the lower end of the distribution; the "tail" is to the right. Also, the mean is to the right of the median, and the mode is to the left of the median

MEASURES OF CENTRAL TENDENCY

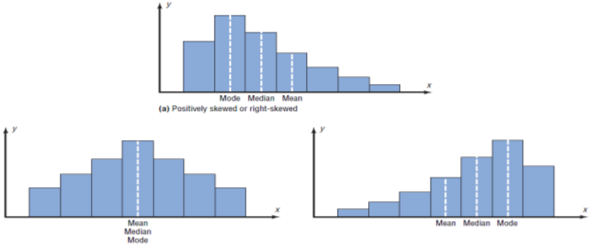
- In a **symmetric distribution** the data values are evenly distributed on both sides of the mean. In addition, when the distribution is unimodal, the mean, median, and mode are the same and are at the center of the distribution.

- When the majority of the data values fall to the right of the mean and cluster at the upper end of the distribution, with the tail to the left, the distribution is said to be **negatively skewed or left-skewed**

MEASURES OF CENTRAL TENDENCY

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** $\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$
- **Sample variance** $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$
- Population Standard deviation $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$
- **Sample standard deviation** $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

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 variation**
- The **coefficient of varaiation**, denoted by CVar, is the standard deviation divided by the mean. The result is expressed in percentage
- **For Samples**, CVar $= \frac{s}{\bar{x}} * 100$
- **For Samples**, CVar $= \frac{\sigma}{\mu} * 100$

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 $= \frac{s}{\bar{x}} = \frac{5}{87} * 100 = 5.7\%$ sales.

CVar $= \frac{773}{5225} * 100 = 14.8\%$ commissions.

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

MEASURES OF VARIATION

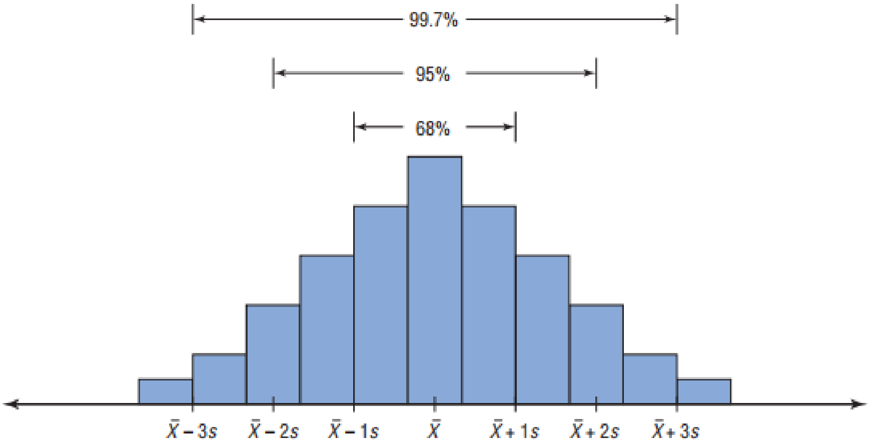
- **The Empirical (Normal) Rule** – Chebyshev's theorem applies to any distribution regardless of its shape. However, when a distribution is bell-shaped (or what is called normal), the following statements, which make up the empirical rule, are true

- 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 and 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 $D_1, D_2, \ldots D_{99}$
- Percentiles – divide an ordered data set into approximately 100 equal parts. $P_1, P_2, \ldots P_{99}$

Compute Q_1, Q_2 and Q_3 and find the inter-quartile range of the data:

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: $6\,,7\,,8\,,\mathbf{10}\,,11\,,15\,,17\,,\mathbf{18}\,,18\,,19\,,20\,,\mathbf{31}\,,54\,,59\,,104$
   $Q_2 = 18\,, Q_1 = 10$ and $Q_3 = 31$

   $\text{IQR} = Q_3 - Q_1 = 31 - 10 = 21$
2. Ordered data : $3\,,6\,,\mathbf{7}\,,\mathbf{8}\,,12\,,15$

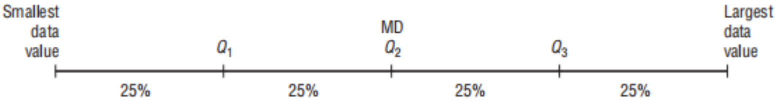
   $Q_2 = \frac{7+8}{2} = 7.5\,, Q_1 = 6$ and $Q_3 = 12$.
   $\text{IQR} = Q_3 - Q_1 = 12 - 6 = 6$
3. Ordered data: 4 , 15 , 19 , **20** , 25 , 33 , 37
   $Q_2 = 20\,, Q_1 = 15$ and $Q_3 = 33$
   $\text{IQR} = Q_3 - Q_1 = 33 - 15 = 18$

MEASURES OF POSITION

Quartiles and Deciles

Quartiles divide the distribution into four groups, separated by Q_1, Q_2, Q_3.

Note that Q_1 is the same as the 25th percentile; Q_2 is the same as 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

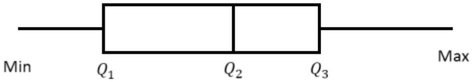
Check the the data set for outliers. 5, 6, 12, 13, 15, 18, 22, 50

BOX PLOT

Box plot

Requires five-number summary namely:

- Minimum entry
- First quartile, Q_1
- Second quartile (or median), Q_2
- Third quartile, Q_3
- Maximum entry



BOX PLOT

Information Obtained from a Boxplot

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, 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

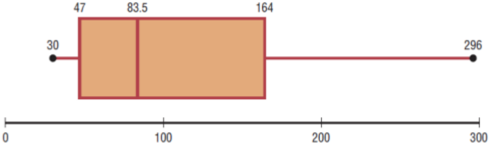
QUESTION

- 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 $Max = 296$;

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



The distribution is somewhat positively skewed.

INTRODUCTION TO PROBABILITY

- 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 even} $= \{2\,,4\,,6\}$

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

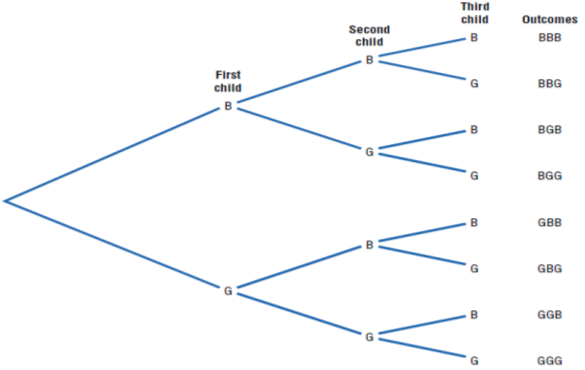
- Find the sample space for rolling two dice
- Find the sample space for tossing a coin thrice

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

TREE DIAGRAM

TREE 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.

  $$P(E) = \frac{Number\,of\,outcomes\,in\,event\,E}{Number\,of\,outcomes\,in\,sample\,space} = \frac{n(E)}{S}$$

- **Empirical** Probability – relative frequency of an event.

  $$P(E) = \frac{Frequency\,of\,Event\,E}{Total\,frequency} = \frac{f}{n}$$

- **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

<u>Solution</u>

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

PROBABILITY

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 denote 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

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

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

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 $\qquad$ Y = respondent answered yes

F = respondent was a female $\qquad$ 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 $P(M|N)$.

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

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

Probablity

- **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)
- $P(B/A) = P(B)$ or $P(A/B) = P(A)$
- Events that are not independent are **dependent**.

PROPABILITY

QUESTION

- 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

PROPABILITY

The Multiplication Rule

- The probability that two events A and B will occur in sequence is
- P(A and B) $= P(A \cap B) =$ 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

<u>Solution</u>

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

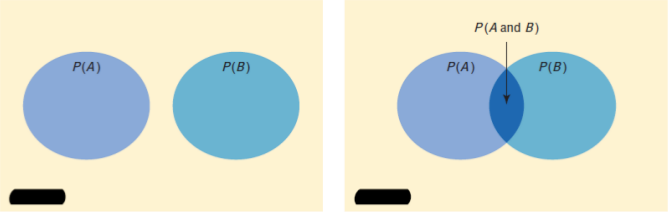
An urn contains 3 red balls, 2 blue balls, and 5 white balls. A ball is selected and its colour noted, then it is replaced. A second ball is selected and its colour noted. Find the probability of each of these.

- 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 occur is

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

$S = \{1, 2, 3, 4, 5, 6\}; \;\; A = \{1, 2\}; \; B = \{1, 3, 5\}; \;\; A \cap B = \{1\}$

- $P(A \; or \; B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- $P(A \cup B) = \frac{2}{6} + \frac{3}{6} - \frac{\mathbf{1}}{\mathbf{6}} = \frac{\mathbf{2}}{\mathbf{3}}$.

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 $\boldsymbol{O}$ or type $\boldsymbol{A}$ blood.

b) type $\boldsymbol{B}$ or is $\boldsymbol{Rh}$ **negative**.

ADDITION RULE

PROBABILITY

Solution

a) $P(O \text{ or } A) = P(O) + P(A) = \frac{184}{409} + \frac{164}{409} = \frac{348}{409}$.

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, there 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 Democrat 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

QUESTION

Example 4: In a 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

$$P(N \text{ or } M) = P(N) + P(M) - P(N \text{ and } M).$$
$$= \frac{8}{13} + \frac{3}{13} - \frac{1}{13} = \frac{10}{13}.$$

ADDITION RULE

PROBABILITY

<u>Test 1</u>

1. If A and B are independent events with $P(A) = 0.3$ and $P(B) = 0.6$, 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 and Counting

- **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 is $\boldsymbol{m} \times \boldsymbol{n}$.

- Can be extended for any number of events occurring in sequence.

Example: You are purchasing a new car. The possible 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 manufacturer, one car size and one colour?

PROBABILITY AND COUNTING

Solution

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

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

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
- $_nP_r = \dfrac{n\,!}{(n-r)\,!}$; where $r \leq n$.

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

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

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

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

PROBABILITY

QUESTIONS

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

Ans: $n = 10$ and $r = 4$.

$$_{10}P_4 \;=\; \frac{10\,!}{(10-4)\,!} = 10 \times 9 \times 8 \times 7 = 5{,}040 \text{ ways.}$$

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 $r = 5$.

$$_5P_5 \;=\; \frac{5\,!}{(5-5)\,!} = 5! = 120 \text{ ways.}$$

COUNTING TECHNIQUES

Distinguishable Permutations

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

Example 5: In how many ways can the letters of the word $\boldsymbol{MISSISSIPPI}$ be arranged?

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

Example 6: In how many ways can the letters of the word $\boldsymbol{STATISTICS}$ be arranged?

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

COUNTING TECHNIQUES

Combinations

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

$$10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 10^7 = 10{,}000{,}000$$

possible identification numbers

- Only one of those numbers corresponds to your ID number

$$\therefore P(your\ ID\ number) = \frac{1}{10{,}000{,}000}.$$

QUESTIONS

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 $\dfrac{11\,!}{1! \times 4\,! \times 4\,! \times 2\,!} = 34{,}650$

distinguishable permutations of the given letters.

$\therefore P(Mississippi) = \dfrac{1}{34650} = 0.000029.$

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.

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

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

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

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

b) Probability of more women than men

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

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

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

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

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

SOLUTIONS

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

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

QUESTIONS

Example 6: A food manufacturer is analyzing a sample of 100 cashew nuts for the presence of a toxin. In this sample, three nuts have dangerously high levels of the 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

$$P(1\ toxic\ nut) = \frac{{}_{3}C_{1} \cdot {}_{97}C_{3}}{{}_{100}C_{4}} = 0.1128\,.$$

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

TRY THESE

- Find the number of ways can the 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

TRY THESE

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 the 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

PROBABILITY DISTRIBUTION

- Solution

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

- 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

SOLUTIONS

- Construct a probability distribution for the number of times a head shows up when a coin is tossed twice.

Let X represent the number heads that show up.

$S = \{HH,\ HT,\ TH,\ TT\}$

$X = 0, 1, 2.$

$P(X = 0) = P(TT) = \frac{1}{4} = 0.25$

$P(X = 1) = P(HT \text{ or } TH) = \frac{2}{4} = 0.50$

$P(X = 2) = P(HH) = \frac{1}{4} = 0.25$

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. $\sum P(X) = 1$

<u>Example</u>

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

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

i. Find the value of a.

ii. Find $\mathrm{P}(X \leq 2)$.

RANDOM VARIABLES

Solution

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

$0.15 + 0.35 + a + 0.10 + 0.05 + 0.20 = 1$

$a = 1 - 0.85 = 0.15.$

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

$0.15 + 0.35 + 0.15 = 0.65.$

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

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

$\sigma^2 = 0.5(0-0.5)^2 + 0.5(1-0.5)^2 = 0.25$

$\sigma = \sqrt{0.25} = 0.50$

RANDOM VARIABLES

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

Solution

$S = \{BB, BG, GB, GG\}$; thus $X = 0, 1, 2.$

$P(X = 0) = \frac{1}{4} = 0.25;\ P(X = 1) = \frac{2}{4} = 0.50;\ P(X = 2) = \frac{1}{4} = 0.25$

X	0	1	2
$P(X)$	0.25	0.50	0.25

$$\mu = 0(0.25) + 1(0.50) + 2(0.25) = 1$$

$$\sigma^2 = 0.25(0-1)^2 + 0.50(1-1)^2 + 0.25(2-1)^2 = 0.50$$

$$\sigma = \sqrt{0.50} = 0.7071$$

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

1. 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)	$\frac{1}{1000}$	$\frac{999}{1000}$

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

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 $\mathrm{P}(X)$	$\frac{2}{1000}$	$\frac{2}{1000}$	$\frac{2}{1000}$	$\frac{2}{1000}$	$\frac{992}{1000}$

$$
\begin{aligned}
E(X) &= \$98\left(\frac{2}{1000}\right) + 48\left(\frac{2}{1000}\right) + 23\left(\frac{2}{1000}\right) + 8\left(\frac{2}{1000}\right) + (-\$2)\left(\frac{992}{1000}\right) \\
&= -\$1.63
\end{aligned}
$$

RANDOM VARIABLES

<u>Test 2</u>

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

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

   a) Find the value of y.
   b) Find $\mathrm{E}(X)$
   c) Find $\mathrm{Var}(X)$
   d) Find $\mathrm{SD}(X)$

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

2020/2021

Department of Statistics and Actuarial Science, KNUST