Statistics and Probability

First Year

Computer Science

Prepared by

Dr/ Marwa Yahya

2025

رؤيــة ورســالة الأكاديمية الحديثة

- **الرؤيــة :**

تتطلع الأكاديمية الحديثه لعلوم الكمبيوتر وتكنولوجيا الإدارة إلى أن تكون متميزة فى مجالات تخصصاتها لمسايرة المستجدات المحلية والإقليمية والعالمية فى سوق الأعمال.

- **الرسالة :**

تلتزم الأكاديمية الحديثه لعلوم الكمبيوتر وتكنولوجيا الإدارة بإعداد كوادر متخصصة فى مجالات علوم الحاسب الآلى وإدارة الأعمال والمحاسبة والمراجعة واقتصاديات التجارة الدولية ونظم معلومات الأعمال وذلك لإمداد المجتمع المحلى والعربى بالكوادر البشرية المؤهلة والمزودة بالأسس النظرية والتطبيقية اللازمة لسوق العمل فى التخصصات المذكورة، ومن خلال الاستفادة من جميع الموارد المتاحة تواكب الأكاديمية التطورات العلمية والتكنولوجية بأنشطتها البحثية كما تساهم فى خدمة المجتمع وتنمية البيئة فى محيطها، ويتم ذلك فى إطار من الإلتزام بالقيم الأخلاقية والعلمية المتعارف عليها.

Vision:

Achieving excellence in its fields of specializations to be in line with the local, regional, and international updates in the labor market.

Mission

Modern academy is committed to preparing professional graduates specialized in the fields of Computer Science, Business Administration, Accounting and Auditing, Economics of International Trade and Business Information Systems to provide the regional and Arab community with qualified cadres equipped with theoretical and professional foundations required in the labor market in the aforementioned fields. By utilizing all the available resources, the academy keeps up with the scientific and technological advancements through research activities in addition to participating in serving the society and developing the surrounding environment, all this within a frame of commitment to the recognized moral and scientific values.

رؤية برنامج علوم الحاسب

تحقيق التميز في مجال علوم الحاسب محليا و إقليميا و دوليا، في ضوء المعايير المعتمدة لجودة التعليم.

- Computer Science Program Vision

Achieving excellence in the field of Computer Science locally, regionally and internationally, in view of the approved standards of the quality of education.

رسالة برنامج علوم الحاسب

يلتزم برنامج علوم الحاسب بالأكاديمية الحديثة لعلوم الكمبيوتر و تكنولوجيا الإدارة بتقديم خدمات تعليمية مطورة تواكب معايير جودة التعليم بما يسهم في إعداد خريج متميز له القدرة على المنافسة في تخصص علوم الحاسب, و لديه القدرة على إجراء أبحاث علمية متقدمة و تقديم خدمات فعالة للمجتمع و البيئة المحيطة .

- Computer Science Program Mission

The Computer Science program in the Academy is committed to providing updated educational services that match the standards of the quality of education, in order to prepare a distinguished graduate having the ability to compete in the field of computer science, conduct advanced scientific researches and provide effective services to the society and surrounding environment.

Modern Academy for Computer Science
And Management Technology in Maadi
Department: Basic Sciences

Course Specifications

1- Basic Information	
Academic year / Level: 1st year /2nd term	**Specialization: Computer Science**
Title: Statistical and Probabilities	**Code: - BS110**
Lecture: 2 Tutorial: 2 Practical: ---- Total: 3 (Hour/week)	
2 – Overall Aims of Course:	**On completion of this course the successful student will be able to:** Enable graduates to exhibit a high level of practical and theoretical skills in Mathematics with knowledge of currently available techniques and technologies. Explore the principles that support developments in Mathematics. Teach students basic mechanisms for following and learning the continuous progress in Mathematics.
3 – Intended Learning Outcomes of Course (ILOs):	
A-Knowledge and Understanding :	**On completion of this course student will have knowledge and understanding of:** a1.Recognizes the concepts of Probability.[A13] a2 - Explain mechanisms and method Probability.[A1] a3 - Demonstrate statistical methods to computer science applications.[A2]
B-Intellectual Skills:	**On completion of this course the successful student will be able to:** b1 – Realize various type of data in statistics and using the appropriate method of Solving the related statistical model.[B3]
C-Professional and Practical Skills:	**On completion of this course the successful student will be able to:** c1- solve different exercises of Probability.**[C4]** c2 – Explain how to treat a computer model as a statistical model.[C6]
D-General and Transferable Skills:	**On completion of this course the successful student will be able to:** d1. Search for other Methodologies in specific topics related to Probability.**[D1]** d2 – Learn how to transform the solution of Statistical model to the solution of desired computer model.[D7]

4-Contents:

No	Contents
1	Probability
2	Some probability distributions
3	Sampling distributions
4	Estimation of the mean and proportion
5	Regression and correlation

5–Teaching and Learning Methods:

Practical training: (X)	- Exercises: (√)	Lectures: (√)
Presentation: (√)	- Projects: (X)	-Open Discussion: (√)
E. Learning: (√)	Self Studies: (√)	-Web-Site searches: (√)
Case Study: (X)	Virtual class (X)	- Chat Room (√)
Virtual lab (X)	Movie Lectures (√)	- Voice Lectures (√)
Simulation lab (X)	- Others (list): (X)	

6-Student Assessment Methods:

A-Student Assessment Methods:

- Assignments. (√)	- Quizzes (√)	- Reports (X)
- Researches (√)	- Projects (X)	- Discussions (√)
-Presentations (X)	- Participation (X)	- Midterm Exam (√)
- Practical Exam. (X)	- Open Book Exam (X)	- Oral Exam (X)
- Final Exam. (√)	Others (list)	

B-Assessment Schedule:

Assessment method	Week no.
1. Assignments	Every week from 2 to 14
2-Quizzes	Weeks 4 and 11
3- Discussions	During semester
4- Researches	During semester
5- Midterm written exam	Week 7
6- Final exam	Week 16

C-Weighting of Assessments:

Assessment Method		Marks	Percentage (%)
1-Year work	1- Assignments	20	20%
	2-Quiz		
	3- Discussions		
	4- Participations		
	6-Midterm Exam	20	20
2-Final Exam		60	60%
Total		**100**	**100%**

7-List of References:

A-lecture notes.

Lecture Notes, "Probability", staff members, **Modern Academy for Computer Science and Management Technology.**

B- Essential books (text books)

[1]Probability and statistics for Engineers and Scientists, Anthony Hayter -4th edition,2006.
[1] Howard Anton, Elementary linear algebra, John Wiley & Sons, Inc., 2013.
[2] David C. Lay, Linear algebra and its application (3rd Edition) Addison Wesley, 2002.
[3]Probability and statistics for Engineers and Scientists,Anthony Hayter –Nelson Education,2012.
[4]Ross, Sheldon. "Probability and statistics foe engineers and scientist". Vol.16, no. m2. Elsevier, New Delhi, 2009.

C- Recommended Books

……………………………………………………………………………………

D- Periodicals, Web-Sites, etc….

…………………………………………………………………………………..

9- Facilities and teaching materials:

1- White Board **Lecture√..Class√ Lab**	√	**2-PC / Laptop** **Lecture√.. Class….. Lab**	√
3- Printers **Lecture…..Class….. Lab…**		**4- Data Show** **Lecture√ .Class Lab**	√
5- White Board for Presentation	√	**6- Laser Pointer**	...
7- Laboratories(List): **Computer labs** Virtual lab Simulation lab	-	**8- Software Packages (list):**	--
9- Library	...	**10- E. Library** **(i,e. Ebsco, IEEE, Egyptian knowledge Bank, etc)**	√
11-Webinars:	√	**12-Chat Room:**	√

	Moodle √ Zoom √ YouTube **WebEx MS'Teams √** **Others(list):**		**What'sUp √** **FaceBook Messenger √** **Imo Viber line skype** **Telegram Others(list)**		
	13- Social Media Networks: **Facebook √ LinkedIn Others(list):**	√	**14- Website/ Mobile Application**		
	15- Internet connection **(Cable or Wireless)**	√	**16- Recording movie System** **(Camera, mic, speaker, etc)**	√	
	17- Supplies and raw materials (list)		**18- Others(list):**		

- Course Coordinator: Dr. Marwa Yehia Signature**:** Marwa Yehia

- Head of Department: Associated Prof, . Hala Meqdad Signature**:** . Hala Meqdad

Date: 1 / 1 / **2025**

Course Intended Learning Outcomes

Year: **2024-2025**	**Academic term : 2nd term**
Title: Statistical and Probabilities	**Code: BS101**
Academic year / level: 1st year	**Specialization: CS**

Course Content	Hours		a. Knowledge and Understanding			b.Intellectual Skills	c.Professional Skills		d.General Skills	
	Lec.	Tut.	a1	a2	a3	b1	c1	c2	d1	d2
Probability	4	2	√			√			√	
Some probability distributions	4	2	√		√		√		√	√
Sampling distributions	6	4		√	√	√	√		√	
Estimation of the mean and proportion	8	2		√	√	√	√	√	√	√
Regression and correlation	6	2	√			√		√	√	√

Teaching and learning methods versus Intended Learning Outcomes

Year: **2024-2025**	**Academic term : 2nd term**
Title: Statistical and Probabilities	**Code: BS101**
Academic year / level: 1st year	**Specialization: CS**

Teaching Activities	**a.Knowledge &Understanding**			**b.Intellectual Skills**	**c.Professional Skills**		**d.General Skills**	
	a1	**a2**	**a3**	**b1**	**c1**	**c2**	**d1**	**d2**
Lectures	√		√		√	√	√	
Exercises	√	√	√	√	√	√	√	√
Open Discussion	√	√	√	√	√	√		
Self Studies	√	√	√	√	√	√	√	√
Presentation	√	√	√		√	√		
Web-site search				√			√	√
E.Learning	√	√	√	√	√	√	√	√
Movie Lecture	√	√	√		√	√	√	
Voice Lecture	√	√	√		√	√	√	
Chat Room	√	√	√	√	√	√	√	√

Course Assessment Methods versus Intended Learning Outcomes

Year: **2024-2025**	**Academic term : 2nd term**
Title: Statistical and Probabilities	**Code: BS101**
Academic year / level: 1st year	**Specialization: CS**

Methods Of Evaluating ILO's	**a.Knowledge &Understanding**			**b.Intellectual Skills**	**c.Professional Skills**		**d.General Skills**	
	a1	**a2**	**a3**	**b1**	**c1**	**c2**	**d1**	**d2**
1 – Assignments	√	√	√	√	√	√	√	√
2- Written Exam (Final+ midterm)	√	√	√	√	√	√		
3-Quizes	√	√	√	√				
4- Participations	√	√	√	√	√	√	√	√
5-Discussions	√	√	√	√	√	√	√	√

Introduction

Statistics simply means numerical data, and is field of math that generally deals with collection of data, tabulation, and interpretation of numerical data. It is actually a form of mathematics analysis that uses different quantitative models to produce a set of experimental data or studies of real life. It is an area of applied mathematics concern with data collection analysis, interpretation, and presentation.

Statistics deals with how data can be used to solve complex problems. Some people consider statistics to be a distinct mathematical science rather than a branch of mathematics.

Statistics makes work easy and simple and provides a clear and clean picture of work you do on a regular basis.

Contents

Chapter 1 PROBABILITY

Chapter 2 Some Probability Distributions

Chapter 3 Sampling Distributions

Chapter 4 Estimation of the mean and proportion

Chapter 5 Regression and Correlation

	CHAPTER
	ONE

PROBABILTY

1.1. Events and the Sample Space

1.2. Simple and Compound Events

1.3. Calculating Probability

1.4. Marginal and Conditional Probability

1.5. Mutually Exclusive Events

1.6. Some Rules for Computing Probabilities

1.7. Bayes' Theorem

PROBABILTY

1.1. Experiment, Outcome, and Sample Space

An experiment is a process that, when performed, result in one and only one of many observations. These observations are called the ***outcomes*** of the experiment.

Experiment

An ***experiment*** is any process of observation that has an uncertain outcome.

Event

An ***event*** is an experimental outcome that may or may not occur.

The collection of the all outcomes for an experiment is called a ***sample space***. A sample space is denoted by S. the sample space for the example of inspecting a tennis ball is written as

$$S = \{good, defective\}.$$

The elements of a sample space are called ***sample points*.**

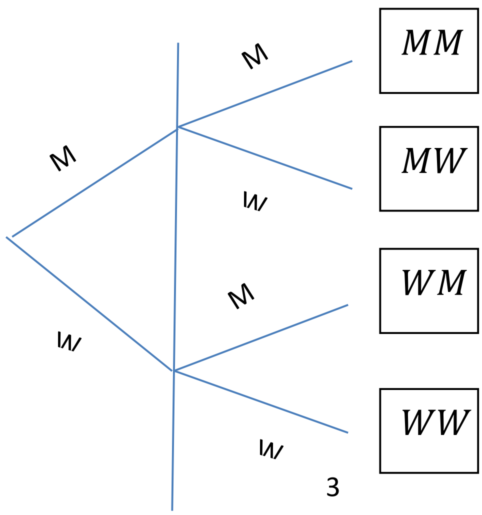
Example (1)

Suppose we randomly select two employees from a company and observe whether the employee selected each time is a man or a woman. Write all the outcomes for this experiment. Draw the tree diagrams for this experiment.

Solution

Let us denote the selection of a man by M and that of a woman by W. There are four final outcomes: *MM*, *MW*, *WM*, *and WW*. Hence, the sample space is written as

$S = \{ MM, MW, WM, WW \}$.



1.2. SIMPLE AND COMPOUND EVENTS

An event may be a ***simple event*** or a ***compound event***. We denote an event by using a capital letter. If the letter E denotes a particular event, then $P(E)$ denotes the probability that the event E will occur.

Event: is a collection of one or more of the outcomes for an experiment.

Simple Event

An event that include one and only one of the (final) outcomes for an experiment is called a simple event and is usually denoted by E_i.

Compound Event

A compound event is a collection of more than one outcome for an experiment.

Example (2)

In a group of people, some are in favor of free trade and others are against it. Two persons are selected at random from this group and asked whether they are in favor or against free trade.

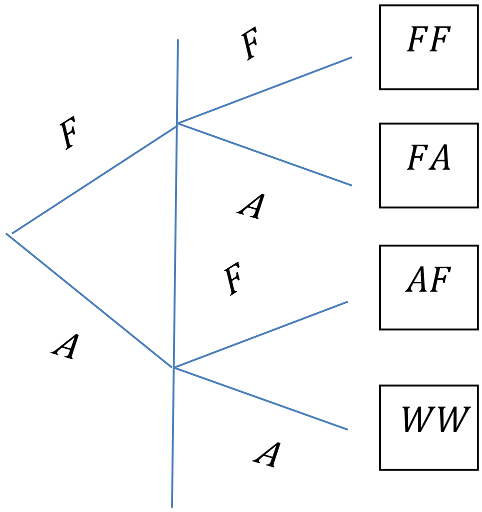
1. How many distinct outcomes are possible?
2. Draw a tree diagram for this experiment.
3. List all the outcomes included in each of the following events and mention whether they are simple or compound events.
    a. Both persons are in favor of free trade.
    b. At most one person is against free trade.
    c. Exactly one person is in favor of free trade.

Solution

Let F = a persons is in favor of free trade.

A = a person is against free trade.

$S = \{FF, FA, AF, AA\}$,



a. Both persons are in favor of free trade = {FF}. Because this event includes only one of the final four outcomes, it is a simple event.
b. At most one person is against free trade = {FF, FA, AF}. Because this event includes more than one outcome, it is a compound event.
c. Exactly one person is in favor of free trade = {FA, AF}. Because this event includes more than one outcome, it is a compound event.

1.3. CALCULATING PROBABILITY

The probability of an event is a number that measures the chance, or likelihood, that the event will occur when the experiment is carried out. The following are two important properties of probability.

1. The probability of an event is greater than or equal to zero and is less than or equal to one.

For an impossible event M *"the event M never occurs"*: $P(M) = 0$.

For a sure event C "the event C is certain to occur": $P(C) = 1$.

Intuitively, then, the closer that $P(E)$ is to one, the higher is the likelihood that the event A will occur, and the closer that $P(E)$ is to zero, the smaller is the likelihood that the event E will occur.

2. **The sum of the probabilities of all simple events (or final outcomes) for an experiment, denoted by $\sum \boldsymbol{P}(\boldsymbol{E_i})$, is always 1.**

Thus, for an experiment

$$\sum P(E_i) = P(E_1) + P(E_2) + P(E_3) + \cdots = 1$$

From this property, for the experiment of one toss of a coin

$$P\ (H)\ +\ P\ (T)\ = 1$$

For the experiment of two tosses of a coin

$$P\ (HH)\ + P\ (HT)\ +\ P\ (TH)\ +\ P(TT)\ =\ 1$$

CLASSICAL PROBABILITY

Outcomes that have the same probability of occurrence are called equally likely outcomes. The classical probability rule is applied to compute the probabilities of events for an experiment all of whose outcomes are equally likely.

CLASSICAL PROBABILITY RULE

$$P(E_i) = \frac{number\ of\ favorable\ outcomes}{Total\ number\ of\ outcomes\ for\ the\ experiment},$$

Example (3)

Find the probability of obtaining a head and the probability of obtaining a tail for one toss of a coin

$$S = \{H, T\}$$

$$P(head) = \frac{number\ of\ favorable\ outcomes}{Total\ number\ of\ outcomes} = \frac{1}{2}$$

Example (4)

Find the probability of obtaining an even number in one roll of a die.

Solution

Let A be an even number is observed on the die

$$S = \{1,2,3,4,5,6\}$$

$$P(A) = \frac{Number\ of\ outcomes\ included\ in\ A}{Total\ number\ of\ outcomes} = \frac{3}{6}$$

1.4. MARGINAL AND CONDITIONAL PROBABILITY

Marginal Probability is the probability of a single event without consideration of any other event.

Conditional Probability is the probability that an event will occur given that another event has already occurred. If A and B are two events, then the conditional probability of A is written as

$$P(A|B),$$

are read as "the probability of A given that B has already occurred".

Example (5)

A card is drawn at random from a deck of cards. What is the probability that it is a jack, given that it is red?

Solution

$$P(\text{a jack}|\text{a red}) = \frac{P(\text{a jack} \cap \text{a red})}{P(\text{a red})} = \frac{2/52}{26/52} = \frac{2}{26}.$$

Example (6)

The following table gives a two-way classification of all 1000 employees of a large company based on whether they are single or married and whether or not they stocks and bonds.

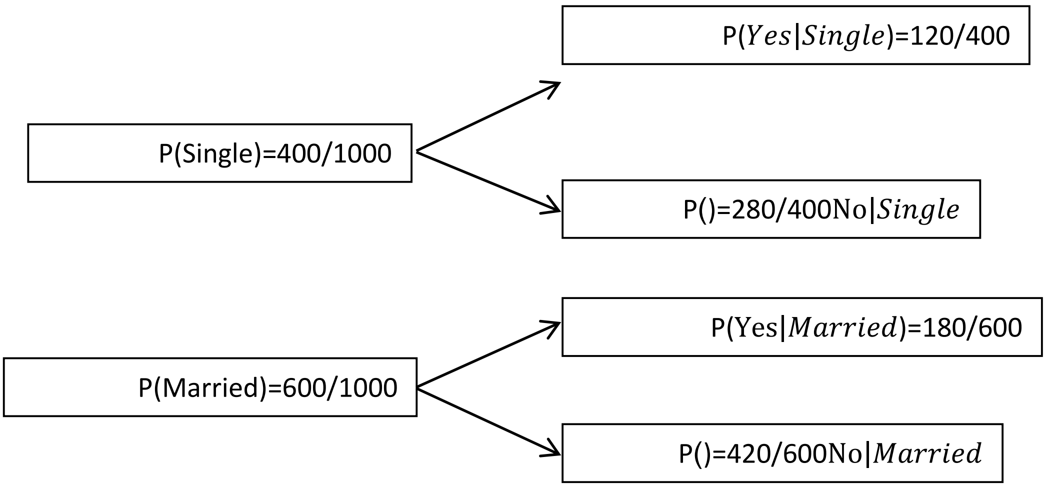
	Own stocks and Bonds	
	Yes	No
Single	120	280
Married	180	420

a. Draw a tree diagram

b. Find all joint probabilities.

Solution

	Own stocks and Bonds		Total
	Yes	No	
Single	120	280	400
Married	180	420	600
Total	300	700	1000



$\mathrm{P(single} \cap \mathrm{yes)} = 120/1000.$

$\mathrm{P(single} \cap \mathrm{no)} = 280/1000.$

$\mathrm{P(married} \cap \mathrm{yes)} = 180/1000.$

$\mathrm{P(married} \cap \mathrm{no)} = 420/1000.$

Example (7)

Five hundred employees were selected from a city's large private companies, and they were asked whether or not they have any retirement benefits provided by their companies. Based on this information, the following two-way classification table was prepared.

	Have Retirement Benefits		Total
	Yes	No	
Men	225	75	300
Women	150	50	200
Total	375	125	500

If one employee is selected at random from these 500 employees, find probability that this employee

a. is a woman.

b. has retirement benefits.

c. Has retirement benefits given the employee is a man.

d. Is a woman given that she does not have retirement benefits.

Solution

a. $\mathrm{P(woman) = 200/500.}$
b. $\mathrm{P(retirement\ benefits) = 375/500.}$
c. $\mathrm{P(retirement\ benefits|man) = 225/300.}$
d. $\mathrm{P(woman|does\ not\ retirement\ benefits) = 50/125.}$

Example (8)

All the 420 employees of a company were asked if they smoke or not and whether they are college graduates or not. Based on this information, the following two-way classification table was prepared.

	College Graduate	Not College Graduate	Total
Smoker	35	80	115
Nonsmoker	130	175	305
Total	165	255	420

a. Draw a tree diagram
b. Find all joint probabilities.

Solution

a.

$$\begin{array}{lcl} & & \boxed{P(\text{college graduate}|\text{Smoker})=165/420} \\ \boxed{P(\text{Smoker})=115} & \begin{array}{c}\nearrow \\ \searrow\end{array} & \\ & & \boxed{P(\text{Not }|\text{Smoker}) =255/420} \end{array}$$

$$\begin{array}{lcl} & & \boxed{P(\text{college graduate}|\text{Nonsmoker})=165/420} \\ \boxed{P(\text{Nonsmoker})=30} & \begin{array}{c}\nearrow \\ \searrow\end{array} & \\ & & \boxed{P(\text{Not }|\text{Nonsmoker})=255/420} \end{array}$$

$$\mathrm{P}(\text{somker} \cap \text{college graduate}) = \frac{35}{420}.$$

$$\mathrm{P}(\text{somker} \cap \text{not college graduate}) = \frac{80}{420}.$$

$$\mathrm{P}(\text{nonsomker} \cap \text{college graduate}) = \frac{130}{420}.$$

$$\mathrm{P}(\text{nonsomker} \cap \text{not college graduate}) = \frac{175}{420}.$$

Example (9)

A quality-control engineer summarized the frequency of the type of defect with the manufacturing of a certain motor. The following table shows which of the three shifts was responsible for the type of defect

Type of Defect					
Shift	Misaligned Component	Missing Component	Measurement Outside of Specification Limits	Other	Total
1	23	13	12	14	62
2	25	15	18	12	70
3	5	11	10	2	28
Total	53	39	40	28	160

a. What is the probability that a defective motor will have a measurement outside of its specification limits?

b. What is the probability that a defective motor was not produce by shift 1?

c. What is the probability that a defective motor produced by shift 3 does not have a misaligned component?

d. What is the probability that a defective motor has a misaligned component or was produced by shift 1?

Solution

a. $P(MOSL) = \frac{40}{160} = 0.25$

b. $(does\ not\ produced\ by\ shift\ 1) = P(Shift\ 2) + P(shift\ 3) = \frac{70}{160} + \frac{28}{160} = 0.6125$

or $(does\ not\ produced\ by\ shift\ 1) = 1 - \frac{62}{160} = 0.6125$

c. $P(does\ not\ Mis \mid shift\ 3) = \frac{P(does\ not\ Mis \cap shift\ 3)}{P(shift\ 3)}$

$$= \frac{(11+10+2)/160}{28/160} = \frac{23}{28} = 0.8214$$

d. $(Mis\ com \cup shift 1) = (mis\ com) +$

Investment Area					
Business	Stocks	Bonds	Commercial Paper	Commodities	Stock Options
Doctors	30	25	15	2	0
Lawyers	29	34	12	0	5
Bankers	50	35	29	5	10
Others	21	14	10	3	2
Total	130	108	66	10	17

$P(shift\ 1) - P(mis\ com \cap shift\ 3) = \frac{53}{160} + \frac{62}{160} - \frac{23}{160} = \frac{92}{160} = 0.575$

Example (10)

An investment newsletter writer wanted to know in which investment areas her subscribes were most interested. A questionnaire was sent to 331 randomly selected professional clients, with the following results

a. What is the probability that an investment client is neither a doctor nor a lawyer?
b. What is the probability that an investment client is a banker and that the investment client's main investment interest is in commodities?
c. If an investment client's main investment interest is commodities, what is the probability that he or she is a banker?
d. What is the probability that an investment client's main investment interest is not in stock options?
e. Let A be the event that an investment client is a lawyer. Let B be the event that an investment client's main investment interest is in commodities. Are the events A and B mutually exclusive?

Solution

a. $\mathrm{P(neither\ a\ doctor\ nor\ a\ lawyer)} = \frac{129}{331} + \frac{50}{331} = \frac{179}{331} = 0.5407.$

<u>Another solution</u>

$\text{P(neither a doctor nor a lawyer)} = 1 - \left(\frac{72}{331} + \frac{80}{331}\right) =$

$1 - \frac{152}{331} = 0.5407.$

b. $\text{P(banker } \cap \text{ commodities)} = 5/331 = 0.0151$

c. $\text{P(banker}|Comm) =$

$P(\text{banker} \cap Comm)/Comm) = \frac{5}{331}/\frac{10}{331} = \frac{5}{10} = 0.5$

d. $\text{P(not stock)} = 1 - \frac{17}{331} = 0.9486$

e. $\text{P(lowyer} \cap \text{commodities)} = 0$, then the two events A and B are mutually exclusive events.

Example (11)

A random sample of American were asked the question, "Do you approve or disapprove of a new Federal tax code that applies one low tax rate to all American?" Assume that the results of 2,000 American are as follows

Political Party	Approve	Disapprove	Don't Know	Total
Democrat	576	313	111	1000
Republican	656	251	93	1000
Total	1232	564	204	2000

Suppose that one of the 2000 American in the survey is selected at random

a. What is the probability that the participant is a Democrat and disapproves?
b. What is the probability that the participant is a Republican or approves?
c. What is the probability that the participant is a Republican or approves or doesn't know?

Solution

a. $\mathrm{P}(\text{Democrat } \cap \text{disapproves}) =$

$313/2000 = 0.1565$

b. $P(\text{Republican } \cup \text{approves }) =$

$$\frac{1000}{2000} + \frac{1232}{2000} - \frac{656}{2000} = \frac{1575}{2000} = 0.788$$

c. $P(\text{Republican } \cup \text{approves } \cup \text{doesn't know}) =$

$$\frac{1000}{2000} + \frac{1232}{2000} + \frac{111}{2000} - \frac{656}{2000} - \frac{93}{2000} = 0.8435$$

1.5. MUTUALLY EXCUSIVE EVENTS

Two events A and B are ***mutually exclusive*** if they have no sample space outcomes in common. Events A and B cannot occur together, and thus

$$P(A \cap B) = 0.$$

Example (12)

In a sample of 460 persons, 120 own stocks, 155 own bonds, and 225 own neither stocks nor bonds. Of the 120 persons who own stocks and 155 who bonds, 40 own both stocks and bonds. Suppose one person is selected at random from these 460 persons. Let

A = the event that a randomly selected person owns stocks.

B = the event that a randomly selected person owns bonds.

C = the event that a randomly selected person owns neither stocks nor bonds.

Are events A and B mutually exclusive? What about events A and C? What about events B and C?

Solution

- First consider events A and B, if the selected person is one of the 40 who own both stocks and bonds, events A and B both happen.

- Now, consider events A and C. these events don not have any common outcomes. Hence, events A and C are mutually exclusive events.
- Similarly, because events B and C do not have any common outcome, they are also mutually exclusive events.

1.6. SOME RULES FOR COMPUTING PROBABITITES

1. The ***intersection of A and B*** is the event consisting of the sample space outcomes belonging to both A and B. the intersection is denoted by $A \cap B$. Furthermore, $P(A \cap B)$ ***denotes the probability that both A and B will simultaneously occur.***
2. The ***union of A and B*** is the event consisting of the sample space outcomes belonging to either A and B. the union is denoted $A \cup B$. Furthermore, $P(A \cup B)$ denotes ***the probability that either A or B will occur.***

General Rule of addition

$$P(A \text{ } or \text{ } B) = P(A) + P(B) - P(A \text{ } and \text{ } B).$$

Special Rule of Addition

$$P(A \text{ } or \text{ } B) = P(A) + P(B).$$

General Rule of Multiplication

$$P(A \text{ and } B) = P(A).P(B|A).$$

Special Rule of Multiplication

$$P(A \text{ and } B) = P(A).P(B).$$

Conditional Probability

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}.$$

Independent Events

Two events are said to be independent if the occurrence of one does not affect the probability of the occurrence of the other. In other words, A and B are independent events if either

$$P(A|B) = P(A) \qquad or \qquad P(B|A) = P(B).$$

Two Important Observations

The following are two important observations about mutually exclusive, independent, and dependent events.

1. Two events are either mutually exclusive or independent. In other word,
   a. Mutually exclusive events are always dependent.
   b. Independent events are never mutually exclusive.
2. Dependent events may or may not be mutually exclusive.

Complementary events

The complement of event A, denoted by $\bar{A}$ and read as "a bar" or " A complement" is the event that includes all the outcomes for an experiment that are not in A.

Example (13)

A card is drawn at random from a deck of cards. What is the probability that

a. A jack
b. A spade
c. A jack or an ace
d. A jack or a spade

Solution

a. $\mathrm{P(a\ jack)} = \frac{4}{52}$.　　　　b. $\mathrm{P(a\ spade)} = \frac{13}{52}$.

c. P(a jack or an ace)

$$= \text{P(a jack)} + \text{P(an ace)} - \text{P(a jack and an ace)}.$$

$$\text{P(a jack or an ace)} = \frac{4}{52} + \frac{4}{52} - \frac{0}{52} = \frac{8}{52}.$$

e. P(a jack or a spade) =

$$\text{P(a jack)} + \text{P(a spade)} - \text{P(a jack and a spade)}$$

$$\text{P(a jack or an ace)} = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}.$$

Example (14)

If you select a card from a deck of cards, what is the probability that the selected card is a queen or a heart.

Solution

$$\mathrm{P(Q \cup H)} = (Q) + P(H) - (Q \cap H) =$$

$$\frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}.$$

Example (15)

In a particular city, 20% of people subscribe to the morning newspaper, 30% subscribe to the evening newspaper, and 10% subscribe to both. Determine the probability that an individual from this city subscribes to the morning newspaper, the evening newspaper, or both.

Solution

$$\boldsymbol{P}(\text{morning newspaper, the evening newspaper, or both})$$
$$= 0.20 + 0.30 - 0.10 = 0.40$$

Example (16)

A card is drawn at random from a deck of cards. What is the probability that it is a jack, given that it is red?

Solution

$$\text{P}(\text{a jack}|\text{a red}) = \frac{\text{P}(\text{a jack} \cap \text{a red})}{\text{P}(\text{a red})} = \frac{2/52}{26/52} = \frac{2}{26}.$$

Example (17)

A small company has 20 employees and 12 of them are married. If 2 employees are randomly selected from this company, what is the probability that the first of them is married and the second is not?

Solution

$$P(\text{the first is married and the second is not}) =$$

$$\frac{12}{20} \times \frac{8}{19} = 0.253$$

Example (18)

The probability that a randomly selected student from a university is a senior is 0.18, a business major is 0.14, and a senior and a business major is 0.04. Find the probability that a student selected at random from this university is a senior or business major.

Solution

$$\mathrm{P}(\text{a senior or business major}) =$$

$$0.18 + 0.14 - 0.04 = 0.28.$$

Example (19)

If you roll two dice. Find the following probabilities

1. The sum = 9
2. The sum = 5 or 11
3. The first die gives an even number and the second die gives an odd number.

4. The first die gives an even number or the second die gives an odd number.
5. Both of the two dice are the same number.
6. If the first die shows an even number, what is the probability that the sum = 8.

Solution

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

1. P(A) $= \frac{4}{36}$
2. P(B or C) = P(B) + P(C) = $\frac{4}{36} + \frac{2}{36} = \frac{6}{36}$
3. P(D and E) = $P(D \cap E) = \frac{9}{36}$
4. P(D or E) = $P(D \cup E) = \frac{18}{36} + \frac{18}{36} - \frac{9}{36} = \frac{27}{36}$
5. P(F) $= \frac{6}{36}$

6. $\mathrm{P}(\mathrm{G}|D)=\frac{P(G\cap D)}{P(D)}=\frac{3}{36}\div\frac{18}{36}=\frac{3}{36}\times\frac{36}{18}=\frac{3}{18}$.

1.7. BAYES' THEOREM

The revised probability of events A_1, after additional information on another event B_1 is obtained, is calculated as follows.

$$P(A_1|B)=\frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1)+P(A_2)P(B|A_2)}$$

Example (20)

$P(A_1)=0.60, P(A_2)=0.40, P(B|A_1)=0.05$, and $P((B|A_2)=0.10$. Use Bayes' theorem to determine $P(A_1|B)$.

Solution

$$P(A_1|B)=\frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1)+P(A_2)P(B|A_2)}$$

$$= \frac{0.6 + 0.05}{(0.6 \times 0.05) + (0.4 \times 0.10)} = \frac{3}{7}$$

Example (21)

The Northeast Bank of Connecticut has two loan officers, Matt Sanger and Alison Terry, who process all loan applications. Of all the loan applications during the past five years, 55% were processed by Mr. Sanger and 45% were processed by Ms. Terry. It is further known that of all the loan applications processed by Mr. Sanger, 30% were rejected and 70% were approved. Of the loan applications processed by Ms. Terry, 20% were rejected and 80% were approved. One loan application is randomly selected from all the applications processed during the past five years and it is observed that this application was rejected. What is the probability this application was processed by Mr. Sanger?

Solution

Let us define the following events.

A_1 = a loan application is processed by Mr. Sanger.

A_2 = a loan application is processed by Ms. Terry.

B = a loan application is rejected.

C = a loan application is approved.

Suppose one application is selected at random. The prior probabilities of A_1 and A_2 are

$$P(A_1) = 0.55 \qquad \text{and} \qquad P(A_2) = 0.45$$

The conditional probabilities

$$P(B|\,A_1) = 0.30, \qquad P(C|\,A_1) = 0.70$$
$$P(B|\,A_2) = 0.20, \qquad P(C|\,A_2) = 0.80$$

The probability that a randomly selected loan application was processed by Mr. Sanger, when it is known that it was rejected, is obtained by using Bay's rule as follow.

$$P(A_1|B) = \frac{P(A_2)P(B|\ A_2)}{P(A_1)P(B|\ A_1) + P(A_2)P(B|\ A_2)}$$

$$P(A_1|B) = \frac{(0.45)(0.20)}{(0.55)(0.30) + (0.45)(0.20)} = 0.353.$$

In a certain northeastern state that is going through financial difficulties, it is believed that 5% of the banks will fail. It is known that the deposits of 90% of the banks in this state are insured by the Federal Depository Insurance Company (FDIC). It is also believed, from past experience, that 3% of the banks protected by FDIC will fail. A bank examiner employed by the federal government would like to know;

a. What is the probability that, for a randomly chosen bank, the bank has deposits protected by FDIC and the bank will fail?
b. What is the probability that, for a randomly chosen bank, the bank has deposits covered by FDIC or the bank will fail?
c. What percentage of the banks that go under have deposits protected by FDIC?

Solution

a. The first step is to define appropriate events:

$$A = bank\ has\ deposits\ protected\ by\ FDIC.$$

$$B = bank\ will\ fail$$

We note translate each of the statements into a probability. We have the following marginal probabilities:

$$P(A) = 0.90, \ \ P(B) = 0.05$$

The last statement in the problem can be written as “Given that a bank accounts protected by FDIC, the probability that the bank will fail is 0.03.” So this a conditional probability, namely,

$$P(B/A) = 0.03.$$

What does equation 1 ask for? $P(A\ or\ B)$? $P(A/B)$? $P(A\ and\ B)$?. The examiner wishes to know the probability that a bank is protected by FDIC and will fail. This is $P(A\ and\ B)$. Using the multiplicative rule

$$P(A\ and\ B) = P(B/A).P(A) = (0.03)(0.90) = 0.027.$$

b. $P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$

$$P(A\ or\ B) = 0.90 + 0.05 - 0.027 = 0.923$$

Thus, 92.3% of the banks are covered by the FDIC, will fail, or both.

c. This question can be phrased as, “given that a bank has failed, what is the probability that this bank has deposits protected by FDIC?” This is

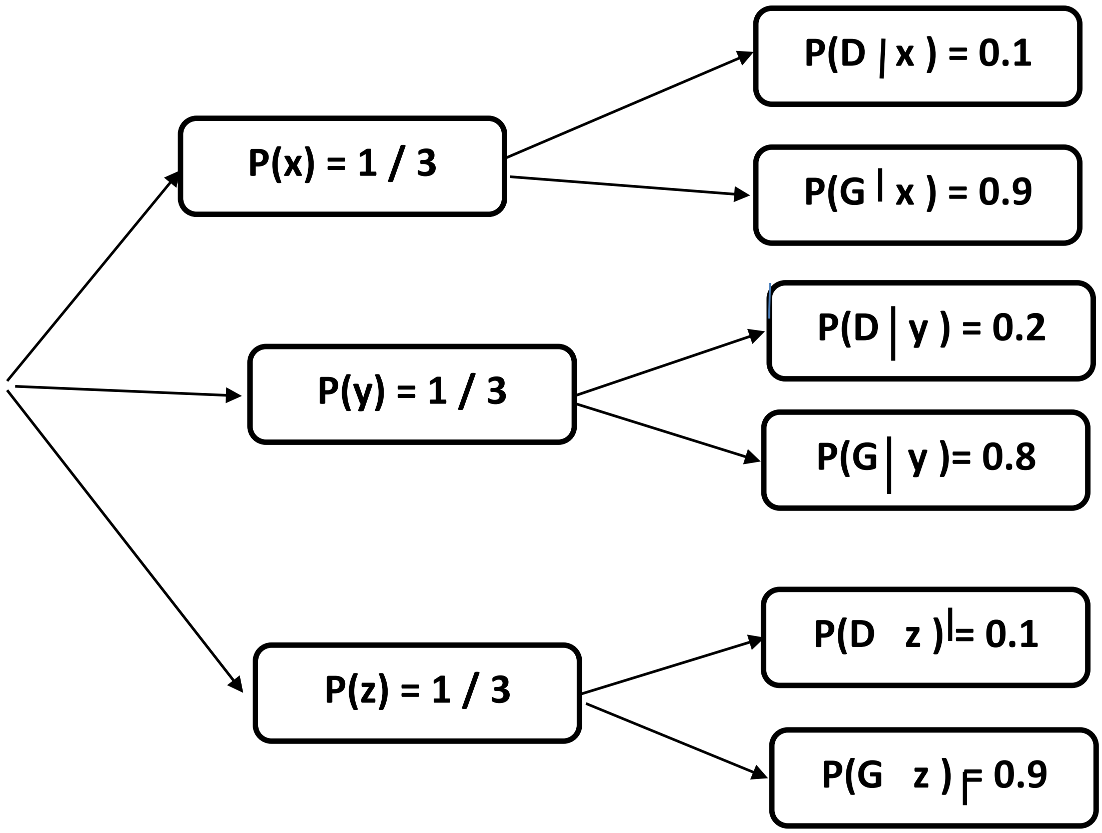
$$P(A/B) = P(A\ and\ B)\backslash P(B) = 0.027\backslash 0.05 = 0.45.$$

Therefore, 54% of those banks that fail have deposits protected by FDIC.

Example (22)

Three machines x, y, and z of equal capacities are producing bullets the probabilities that the machines produce defectives are 0.1, 0.2, and 0.1 respectively. A bullet is taken at random from a day’s production and is found to be defective. What is the probability that it came from machine $\boldsymbol{x}$?

Solution



$$\boldsymbol{P(x|D) = \frac{P(x \cap D)}{P(D)}}$$

$$\boldsymbol{= \frac{P(D|x).P(x)}{P(D|x).P(x) + P(D|y).P(y) + P(D|z).P(z)}}$$

$$\boldsymbol{P(x|D) = \frac{P(x \cap D)}{P(D)}}$$

$$\boldsymbol{= \frac{\left(\frac{1}{3}\right) \times (0.1)}{\left(\frac{1}{3}\right) \times (0.1) + \left(\frac{1}{3}\right) \times (0.2) + \left(\frac{1}{3}\right) \times (0.1)} = 0.25}$$

Example (23)

The contents of urns 1, 2, 3, are as follows

1 white, 2 black, 3 red balls.

2 white, 1 black, 1 red balls.

4 white, 5 black, 3 red balls.

One urn is chosen at random and two balls drawn. They happen to be white and red. What is the probability that they came from urn 2 or 3?

Solution

$$\boldsymbol{P(U1) = 1/3, \; P(U2) = 1/3, P(U3) = 1/3}$$

$$\boldsymbol{P(W \text{ and } R|U1) = \frac{1}{6} \times \frac{3}{5} + \frac{3}{6} \times \frac{1}{5} = \frac{3}{30} + \frac{3}{30} = \frac{6}{30} = \frac{1}{5}}$$

$$\boldsymbol{P(W \text{ and } R|U2) = \frac{2}{4} \times \frac{1}{3} + \frac{1}{4} \times \frac{2}{3} = \frac{4}{12} = \frac{1}{3}}$$

$$\boldsymbol{P(W \text{ and } R|U3) = \frac{4}{12} \times \frac{3}{11} + \frac{3}{12} \times \frac{4}{11} = \frac{2}{11}}$$

$$\boldsymbol{P(U2|A) = \frac{P(U2 \cap A)}{P(A)}}$$

$$\boldsymbol{= \frac{P(U2 \cap A)}{P(U1 \cap A) + P(U2 \cap A) + P(U3 \cap A)}}$$

$$\boldsymbol{P(U2|A)}$$

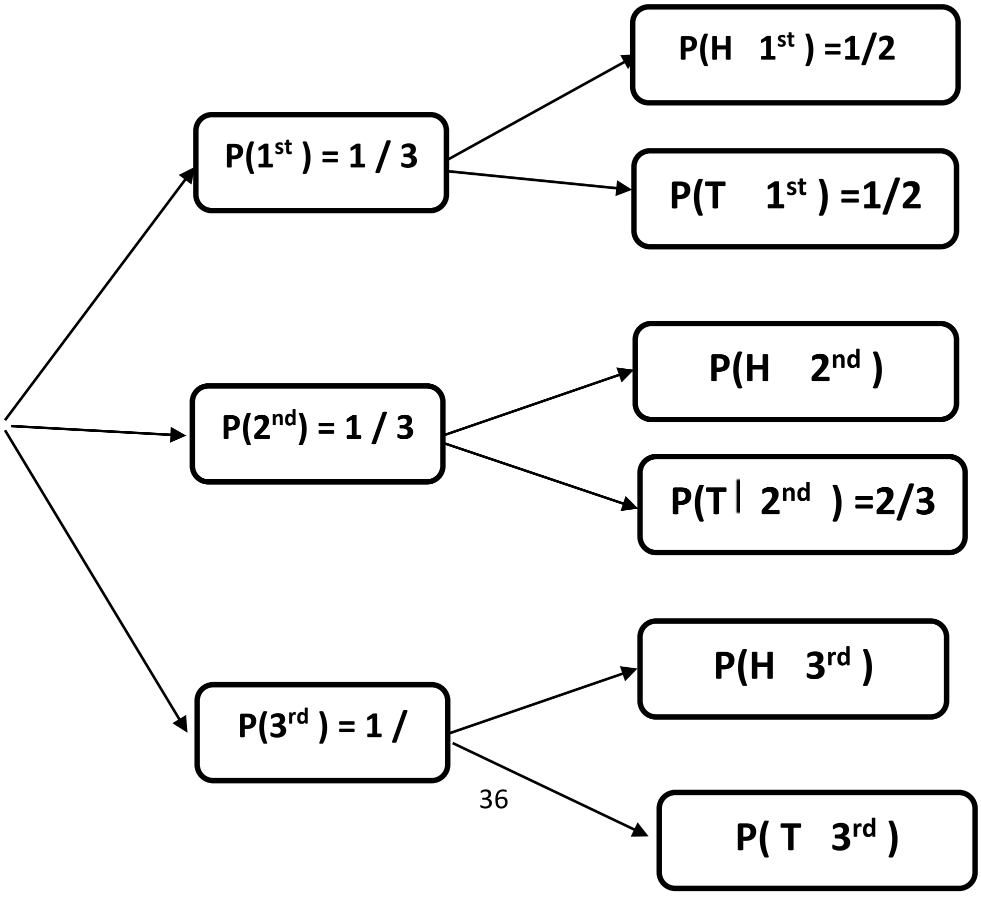
$$\boldsymbol{= \frac{P(A|U2).P(U2)}{P(A|U1).P(U1) + P(A|U2).P(U2) + P(A|U3).P(U3)}}$$

$$P(U2\lfloor A) = \frac{\left(\frac{1}{3}\right)\times\left(\frac{1}{3}\right)}{\left(\frac{1}{5}\right)\times\left(\frac{1}{3}\right)+\left(\frac{1}{3}\right)\times\left(\frac{1}{3}\right)+\left(\frac{2}{11}\right)\times\left(\frac{1}{3}\right)} = \frac{55}{118}$$

Example (24)

There are 3 coins, identical in appearance, one of which ideal and the other two biases with probabilities $\frac{1}{3}$ and $\frac{2}{3}$ respectively for a head. One coin is taken at random and tossed twice. If a head appears both the times, what is the probability that the ideal coin was chosen.

Solution



$$\boldsymbol{P(1st|HH) = \frac{P(1st \cap HH)}{P(HH)}}$$

$$\boldsymbol{= \frac{P(HH|1st).P(1st)}{P(HH|1st).P(1st) + P(HH|2nd).P(2nd) + P(HH|3th).P(3th)}}$$

$$\boldsymbol{= \frac{\left(\frac{1}{3}\right) \times \left(\frac{1}{4}\right)}{\left(\frac{1}{3}\right) \times \left(\frac{1}{4}\right) + \left(\frac{1}{3}\right) \times \left(\frac{1}{9}\right) + \left(\frac{1}{3}\right) \times \left(\frac{4}{9}\right)}}$$

$$\boldsymbol{= \frac{\left(\frac{1}{12}\right)}{\left(\frac{1}{12}\right) + \left(\frac{1}{27}\right) + \left(\frac{4}{27}\right)} = \frac{9}{29}}$$

Exercises

1. Define the following terms: experiment, outcomes.
2. Define the following terms: simple event, compound event.
3. Explain the properties of probability.
4. Describe the impossible and sure event. What is the probability for the occurrence of each of these two events?
5. Explain the difference between the marginal and conditional probabilities of events.
6. What is meant by two mutually exclusive events?
7. Explain the meaning of the intersection of two events.
8. Explain the meaning of union of two events.
9. Consider randomly selecting a card from a standard deck of 52 playing cards. We define the following events:

J = the randomly selected card is a Jack.

Q = the randomly selected card is a Queen.

R = the randomly selected card is a red card (that is, a diamond or heart).

a. Are events J and Q mutually exclusive?
b. Are events J and R mutually exclusive?
c. Are events Q and R mutually exclusive?

10. Suppose an employee is selected at random from a large company. Consider the following two events.

B = the employee selected is a blue-collar worker.

W = the employee selected is a white-collar worker.
Are events B and W mutually exclusive?

11. A die is rolled twice. What is the probability that the sum of the faces is less than 7, given that
    a. The first outcome is a 1.
    b. The first outcome is less than 2.
12. An experiment consists of tossing a single die and observing the number of dots that show on the upper face. Events A, B , and C are defined as follows:
    A: Observe a number less than 5
    B: Observe a number less than or equal 3

C: Observe a number greater than 4

Find the probabilities associated with the events below using either the simple event approach or the rules and definitions from this section.

a. S   b. $A|B$ c. B

d. $A \cap B$ e. $A \cap C$ $A \cup C$

13. All the 420 employees a company were asked if they smoke or not and whether they are college graduates or not. Based on this information, the following two-way classification table was

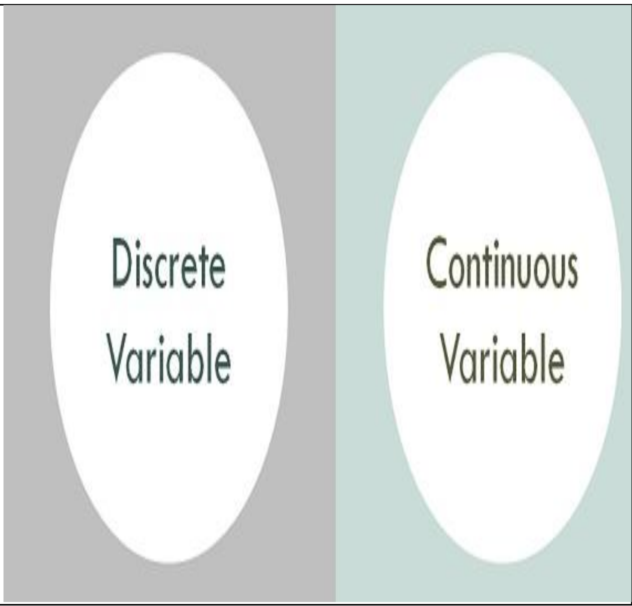
	College Graduate	Not a College Graduate
Smoker	35	80
Nonsmoker	130	175

If one employee is selected at random from this company, find the probability that this employee is a. a college graduate.

b. a nonsmoker

c. a smoker given the employee is a nonsmoker

d. a college graduate given the employee is a nonsmoker
e. P(college graduate and nonsmoker)
f. P(college graduate or smoker)
g. P(smoker or not a college graduate)
h. P(smoker or nonsmoker)



CHAPTER

TWO

Some Probability Distributions

2.1. Discrete Probability Distributions

2.1.1 Probability Distribution of A Discrete Random Variable

2.1.2 Mean and Standard Deviation

2.1.3 Binomial Probability Distribution

2.1.4 Hyper Geometric Distribution

2.1.5 Poisson Probability Distribution

2.2 Continuous Probability Distribution

2.2.1 Normal Probability Distribution

2.2.2 The standard Normal Distribution

Some Probability Distributions

In this chapter the concept of the probability distribution, some probability distributions of a discrete random variable are discussed. Continuous probability distributions, and normal distribution are developed.

2.1 Discrete Probability Distributions

We often use what we called random variable to describe the important aspects of the outcomes of experiments.

Random variable

A ***random variable*** is a variable whose value is determined by the outcome of an experiment. That is, a random variable represents an uncertain outcome. A random variable can be discrete or continuous.

Discrete Random Variable

A random variable that assumes countable values is called a ***discrete random variable***.

Some examples of discrete random variables are:

a. The number of employees working at a company.

b. The number of heads obtained in three tosses of a coin.
c. The number of customers visiting a bank during any given hour.

2.1.1. Probability Distribution of A Discrete Random Variable

The probability distribution of a discrete random variable is a table, a graph, or formula that gives the probability associated with each possible value that the random variable can assume. We denote the probability distribution of the discrete random variable x as $P(x)$.

The probability distribution of a discrete random variable lists all the possible values that the random variable can assume and their corresponding probabilities.

The probability distribution of a discrete random variable $P(x)$ must satisfy the following two conditions.

1. $0 \leq P(x) \leq 1$ for each value of x.
2. $\sum(x) = 1$.

Example (1)

Explain whether or not each of the following is a valid probability distribution. If the probability distribution is valid, show why. Otherwise, show which condition(s) of a probability distribution are not satisfied.

a.

x	P(x)
0.5	-1
0.75	0
1	2

b.

x	P(x)
2	0.25
4	0.35
6	0.30

c.

x	P(x)
0.1	2/7
0.7	4/7
0.9	1/7

Solution

a. This table does not represent a valid probability distribution, because one of the probabilities is negative.

b. This table does not represent a valid probability distribution, because the sum of all probabilities does not equal to 1.0

$$\sum P(x) = 0.25 + 0.35 + 0.30 = 0.90$$

c. Each probability listed in this table is in the range 0 to 1. Also,

$$\sum \mathrm{P(x)} = \frac{2}{7} + \frac{4}{7} + \frac{1}{7} = \frac{7}{7} = 1.$$

Consequently, this table does not represent a valid probability distribution.

Example (2)

The following table gives the probability distribution of a discrete random variable x.

x	0	1	2	3
$P(x)$	0.20	0.15	0.25	0.40

Find the following probabilities

a. $P(x = 1)$.

b. $P(x \leq 1)$.

c. $P(1 < x \leq 3)$.

d. Probability that x assumes a value less than 3.

e. Probability that x assumes a value in the interval 0 to 2.

Solution

a. $P(x = 1) = 0.20.$

b. $P(x \leq 1) = P(x = 0) + P(x = 1) = 0.20 + 0.15 = 0.35.$

c. $P(1 < x \leq 3) = P(x = 2) + P(x = 3) = 0.15 + 0.25 = 0.40.$

d. $P(x < 3) = P(x = 1) + P(x = 2) + P(x = 3) = 0.60.$

e. $P(0 \leq x \leq 2) = P(x = 0) + P(x = 1) + P(x = 2) = 0.80.$

2.1.2 Mean and Standard Deviation

The mean (or expected value) of a discrete random variable, denoted by μ or $E(x)$.

The mean of a discrete random variable x is the value that is expected to occur per repetition, on average, if an experiment is repeated a large number of times. It is denoted by μ and calculated as

$$\mu = \sum x\,P(x),$$

The mean of a discrete random variable x is also called its expected value and denoted by $E(x)$, that is,

$$E(x) = \sum x\,P(x).$$

Example (3)

Let x be the number of errors that a randomly selected page of a book contains. The following table lists the probability distribution of x. Calculate the mean.

x	0	1	2	3	4
$P(x)$	0.73	0.16	0.06	0.04	0.01

Solution

x	$P(x)$	$xP(x)$
0	0.73	0
1	0.16	0.16
2	0.06	0.12
3	0.04	0.12
4	0.01	0.04
$\sum x\,P(x)$		0.44

Taking the mean as the center of a random variable's probability distribution, the ***standard deviation*** is a measure of how much the probability mass is spread out around this center. A higher value for the standard deviation of a discrete random variable indicates that x can assume values over a larger range about the mean. On the other hand, a smaller value for the standard deviation indicates that most of the values that x can assume are clustered closely about the mean. ***The variance*** of a discrete random variable is obtained by squaring its standard deviation.

The standard deviation of a discrete random variable x measures the spread of its probability distribution and is computed as

$$\sigma = \sqrt{\sum x^2 P(x) - \mu^2}$$

Example (4)

Determine the mean, variance, and standard deviation of the random concerning on-the job accidents in the following table

x	0	1	2	3	4
$P(x)$	0.5	0.25	0.10	0.10	0.05

Solution

x	$P(x)$	$x.P(x)$	$x^2.P(x)$
0	0.5	0	0
1	0.25	0.25	0.25
2	0.10	0.20	0.40
3	0.10	0.30	0.90
4	0.05	0.2	0.8
Total	1	0.95	2.35

$$\mu = \sum x.p(x) = 0.95 \text{ accident}$$

and

$$\sigma^2 = \sum x^2.p(x) - \mu^2 = 2.35 - 0.95^2 = 1.45$$

also

$$\sigma = \sqrt{1.45} = 1.20$$

Example (5)

Suppose that a computerized procedure generates integer values between 1 and 5 (that is, 1, 2,3, 4, and 5) in such a way that each value has the same chance of selection. Let X represent a generated value. (i) Describe the probability distribution of X. (ii) What is the mean and variance of this random variable?

Solution

x	$P(x)$	$x.P(x)$	$x^2.P(x)$
0	1\5	0	0
1	1\5	1\5	1\5
2	1\5	2\5	2\5
3	1\5	3\5	9\5
4	1\5	4\5	16\5
5	1\5	1	1\5
Total	5\5=1	3	11

$$\mu = \sum x.p(x) = 3$$

and

$$\sigma^2 = \sum x^2.p(x) - \mu^2 = 11 - 9 = 2$$

also

$$\sigma = \sqrt{2} = 1.414$$

Example (6)

From the following distribution table find the men and the standard deviation.

x	0	1	2	3	4	5
$P(x)$	0.02	0.20	0.30	0.30	0.10	0.08

Solution

x	P(x)	xP(x)	x^2	$x^2P(x^2)$
0	0.02	0	0	0
1	0.20	0.20	1	0.20
2	0.30	0.60	4	1.20
3	0.30	0.90	9	2.70
4	0.10	0.40	16	1.60
5	0.08	0.40	25	2.00
		$\sum xP(x) = 2.50$		$\sum x^2P(x^2) = 7.70$

2.1.3 Binomial Probability Distribution

The ***binomial probability distribution*** is one of the most widely used discrete probability distributions. It is applied to find the probability that an outcome will occur x times in n performances of an experiment. One repetition of a binomial experiment is called ***trial*** or ***Bernoulli trial***.

Binomial experiment

An experiment that consists of n independent, identical trials, each of which results in either a success or a failure and is such that the probability of success on any trial is the same.

Binomial random variable

A random variable that is defined to be the total number of successes in n trials of a binomial experiment.

Binomial distribution

The probability distribution that describes a binomial random variable.

A binomial distribution has the following characteristics:

1. The experiment consists of n trials.
2. Each trial results in a success or a failure.
3. The probability of a success on any trial is p and remains constant from trial to trial. This implies that the probability of failure, q, on any trial is $1 - p$ and remains constant from trial to trial.
4. The trials are independent (that is, the results of the trials have nothing to do with each other).

Furthermore, if x is defined as follows:

x = the total number of successes in n trials of a binomial experiment

The probability of obtaining x success in n trials:

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

where

$n =$ total number of trials.

$p =$ probability of success.

$q = 1 - p =$ probability of failure.

$x =$ number of success in n trials

$n - x =$ number of failures in n trials.

Note that one of the two outcomes of a trial is called a success and the other a failure. Notice that a success does not mean that the corresponding outcome is considered favorable. Similarly, a failure does not necessarily refer to an unfavorable outcome. Success and failure are simply the names used to denote the two possible outcomes of a trial. The outcome to which the question refers is usually called a success; the outcome to which it does not refer is called a failure.

If x is a binomial random variable, then

$$\mu = np \qquad \sigma^2 = npq \qquad \sigma = \sqrt{npq}$$

where n is the number of trials, p is the probability of success on each trial, and $q = 1 - p$ is the probability of failure on each trial.

Example (7)

Consider the experiment consisting of 10 tosses of a coin. Determine if it is a binomial experiment.

Solution

1. There are a total of 10 trials (tosses), and they are all identical. All 10 tosses performed under identical conditions.
2. Each trial (toss) has only two possible outcomes; a head and a tail. Let a head be called a success and a tail be called a failure.

3. The probability of obtaining a head (a success) is 1/2 and that of a tail (a failure) is 1/8 for any toss. That is.

$$P = P(H) = \frac{1}{2} \quad and \quad q = P(T) = \frac{1}{2}$$

The sum of these two probabilities is 1.0. Also, these probabilities remain the same for each toss.

4. The trials (tosses) are independent. The result of any preceding toss has no bearing on the result of any succeeding toss.

Example (8)

Find $P(x = 2)$ for a binomial random variable with $n = 10$ and $P = 0.1$.

Solution

$$P(x = 2) = \frac{10!}{2!\,(10-2)!}(0.1)^2(0.9)^8$$

$$= \frac{10(9)}{2(1)}(0.1)(0.430467) = 0.1937.$$

Example (9)

It is estimated that one out of 10 vouchers examined by audit staff employed by a branch of the Department of Health and Human Services will contain an error. Define X to be the number of vouchers in error out of 20 randomly selected vouchers.

a. What is the probability that at least three vouchers will contain an error?

b. What is the probability that no more than one contains an error?

c. Determine the mean and standard deviation of X?

Solution

$$n = 20, \qquad p = \frac{1}{10}\,(one\ out\ of\ ten)$$

$$P(x) = \frac{n!}{x!\,(n-x)!}p^x(1-p)^{n-x}$$

a.

$$P(X \geq 3) = 1 - P(x < 3)$$

$$= 1 - [P(0) + P(1) + P(2)]$$

$$P(0) = \frac{20!}{0!\,(20-0)!}(0.10)^0(0.90)^{20} = 0.122$$

$$P(1) = \frac{20!}{1!\,(20-1)!}(0.10)^1(0.90)^{19} = 0.270$$

$$P(2) = \frac{20!}{2!\,(20-2)!}(0.10)^2(0.90)^{18} = 0.285$$

$$= 1 - (0.122 + 0.270 + 0.285) = 0.323$$

b.

$$P(X \leq 1) = P(0) + P(1) = 0.122 + 0.270 = 0.392$$

c. The mean of the random variable X is

$$\mu = np = (0.20)(0.10) = 2\, vouchers$$

and the standard deviation of X is

$$\sigma = \sqrt{np(1-p)} = \sqrt{(0.20)(0.10)(0.90)}$$

$$= 1.34\, vouchers$$

This implies that, on the average, the audit staff will encounter 2 vouchers containing an error (out of 20 randomly selected vouchers).

Example (10)

Let the random variable X represent the number of correct responses on a multiple-choice test that has 15 questions. Each question has five multiple choice answers.

a. What is the probability that the random variable X is greater than or equal one if the person taking the test randomly guesses?

b. What is the mean value of X if the person randomly guesses?

c. What is the standard deviation of X if the person randomly guesses?

Solution

$$n = 15, \qquad p = \frac{1}{5} = 0.2$$

$$P(x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

a. $\quad P(x \geq 1) = 1 - P(0)$

$$= 1 - \frac{15!}{0!\,(15-0)!} (0.2)^0 (0.8)^{15}$$

$$= 1 - 0.0351 = 0.9648$$

b. $the\ mean = np = 15(0.2) = 3$

c. $\sigma = \sqrt{np(1-p) = 3(0.8)} = \sqrt{2.4} = 1.5491$

Example (11)

The manager of a retail store knows that 10% of all checks written are “hot” checks. Of the next 25 checks written at the retail store. What is the probability that no more than 3 checks are hot?

Solution

$$n = 25, \;\; p = 0.10$$

$$P(x \le 3) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$$

$$P(x = 0) = \frac{25!}{0!\,(25-0)!}(0.10)^0(0.90)^{25} = 0.0718$$

$$P(x = 1) = \frac{25!}{1!\,(25-1)!}(0.10)^1(0.90)^{24} = 0.1994$$

$$P(x = 2) = \frac{25!}{2!\,(25-2)!}(0.10)^2(0.90)^{23} = 0.2659$$

$$P(x = 3) = \frac{25!}{3!\,(25-3)!}(0.10)^3(0.90)^{22} = 0.2265$$

$$P(x \le 3) = 0.0718 + 0.1994 + 0.2659 + 0.2265$$

$$= 0.7636$$

Example (12)

Over a long period of time, it has been observed that a professional basketball can make a free thrown on a given trial with probability equal to 0.8. Suppose he shoots four free throws.

1. What is the probability that he will make exactly two free throws?
2. What is the probability that he will make at least one free throw?

Solution

1. $P(x = 2) = \frac{4!}{2!2!}(0.80)(0.20) = 0.1536$
2. $P(at\ least\ one) = P(x \geq 1)$

$$= P(1) + P(2) + P(3) + P(4) = 1 - P(0)$$

$$= 1 - \frac{4!}{0!\,4!}(0.8)^0(0.2)^4 = 1 - 0.0016 = 0.9984$$

Although you could calculate $P(x = 1)$, $P(x = 2), P(x = 3)$ and $P(x = 4)$ to find this probability, using the complement of the event makes your job easier; that is

$$P(x \geq 1) = 1 - P(x < 1) = 1 - P(x = 0)$$

2.1.4. Hyper Geometric Distribution

The hyper geometric distribution is a probability distribution that is very similar to the binomial distribution. In section 2.1.3., we learned that one of the conditions required to apply the binomial probability distribution is that the trials are independent so that the probabilities of the two outcomes (success and failure) remain constant. If the trials are not independent, binomial probability distribution cannot be applied to find the probability of x successes in n trials. Such a case occurs when a sample is drawn without replacement from a finite population.

The probability of x successes in n trials is given by

$$P(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

Let

N = total number of elements in the population

r = number of successes in the population.

$N - r$ = number of failure in the population.

n = number of trials.

x = number of successes in n trials.

$n - x$ = number of failures in n trials.

Example (13)

Dawn Corporation has 12 employees who hold managerial position. Of them, 7 are female and 5 are male. The company is planning to send 3 of these 12 managers to a conference. If 3 managers are randomly selected out of 12,

a. find the probability that all 3 of them are female.
b. find the probability that at most 1 of them is female.

Solution

Let the selection of a female be called a success and the selection of a male be called a failure.

a. From the given information

$N = 12$ $r = 7$

$N - r = 5$ $n = 3$

$x = 3$ $n - x = 0$

$$P(x = 3) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} = \frac{\binom{7}{3}\binom{5}{0}}{\binom{12}{3}} = \frac{(35)(1)}{220} = 0.1591$$

b. $N = 12$ $r = 7$

$N - r = 5$ $n = 3$

$x = 0$ $n - x = 3$

$$P(x = 0) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} = \frac{\binom{7}{1}\binom{5}{2}}{\binom{12}{3}} = \frac{(7)(10)}{220} = 0.3182$$

The probability that at most one of the three managers selected is a female:

$$P(x \leq 1) = P(x = 0) + P(x = 1)$$
$$= 0.455 + 0.3182 = 0.3637.$$

2.1.5. Poisson Probability Distribution

The ***Poisson probability distribution***, named after the French mathematician Simeon D. Poisson, is another important probability distribution of a discrete random variable that has a large number of applications. Its probability distribution provides a good model for data that represent the number of occurrences of a specified event in a given unit of time or space.

The following are few examples of experiments for which the random variable x can be modeled by the Poisson random variable:

- The number of calls received by a switchboard during a given period of time.
- The number of accidents that occur at a company during a one-month period.
- The number of television sets sold at a department store during a given week.
- The number of customer arrivals at a checkout counter during a given minute.

The following three conditions must be satisfied to apply the Poisson probability distribution

- x is a discrete random variable.

- The occurrences are random.
- The occurrences are independent.

Let μ be the average number of units that an event occurs in a certain period of time or space. The probability of k occurrences of this event is

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

for values of $k = 0,1,2,3, \dots$. The mean and standard deviation of the Poisson random variable x are

Mean: $\mu,$

Standard deviation: $\sigma = \sqrt{\mu}.$

The symbol $e = 2.71828$ is evaluated using your scientific calculator, which should have a function such as e^x.

Example (14)

Handy Home Center specializes in building materials for home improvements. They recently constructed an information booth in the center of the store. Define X to be the number of customers who arrive at the booth over a 5-minute period. Assume that the conditions for a Poisson situation are satisfied with an average of 4 customers over a 5-minute period.

a. What is the probability that over any 5-minute interval, exactly four people arrive at the information booth?
b. What is the probability that more than two persons will arrive?
c. What is the probability that exactly six people arrive over a 10-minute period?

Solution

a.

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

$$P(4) = \frac{4^4 e^{-4}}{4!} = 0.1954$$

b. $P(X \geq 1) = 1 - P(X < 1) = 1 - [P(0) + P(1)]$

$$P(X \geq 1) = 1 - \left[\frac{4^0 e^{-4}}{0!} + \frac{4^1 e^{-4}}{1!}\right]$$

$$= 1 - [0.0183 + 0.0733] = 0.9084$$

c. $\mu = 4(we\ expect\ four\ people\ over\ a\ 5 - minute\ period)$.

$\mu = 8(we\ expect\ four\ people\ over\ a\ 10 - minute\ period)$.

$$P(6) = \frac{8^6 e^{-8}}{6!} = 0.1221$$

Example (15)

An average of five books per week are returned to a bookstore. Assume that the number of returned books is Poisson distributed.

a. What is the probability that less than two books will be returned in one week?
b. What is the standard deviation of the distributed of the number of books returned in one week?

Solution

a. $P(x < 2) = P(x = 0) + P(x = 1) = \frac{5^0 e^{-5}}{0!} + \frac{5^1 e^{-5}}{1!}$

b. $\mu = \sigma^2 = 5, \quad \sigma = \sqrt{5}$

(Poisson distribution: $\mu = \sigma^2$)

Example (16)

The average number of traffic accident on a certain of highway is two per week. Assume that the number of accidents follows a Poisson distribution with $\mu = 2$.

1. Find the probability of no accidents on this section of highway during a 1-week period.
2. Find the probability of at most three accidents on this section of highway during a 2-week period.

Solution

1. The average number of accidents per week is $\mu = 2$. Therefore, the probability of no accidents on this section of highway during a given week is

$$P(x = 0) = P(0) = \frac{2^0 e^{-2}}{0!} = e^{-2} = 0.135335.$$

2. During a 2-week period, the average number of accidents on this section of highway is 2(2) = 4. The probability of at most three accidents during a 2-week period is

$$P(x \leq 3) = P(0) + P(1) + P(2) + P(3),$$

$$P(x = 0) = P(0) = \frac{4^0 e^{-4}}{0!} = 0.18316$$

$$P(x = 1) = P(0) = \frac{4^1 e^{-4}}{1!} = 0.07326$$

$$P(x = 2) = P(0) = \frac{4^2 e^{-4}}{2!} = 0.14652$$

$$P(x = 3) = P(0) = \frac{4^3 e^{-4}}{3!} = 0.19536$$

$$P(x \leq 3) = 0.18316 + 0.07326 + 0.14652 + 0.19536$$
$$= 0.433471.$$

Example (17)

A washing machine in a Laundromat breaks down an average of three times per month. Using the Poisson probability distribution formula, find the probability that during the next month this machine will have

a. Exactly two breakdowns

b. At most one breakdown.

Solution

a. The probability that exactly two breakdowns will be observed during the next month is

$$P(x = 2) = \frac{3^2 e^{-3}}{2!} = \frac{(9)(0.049787)}{2} = 0.2240$$

b. The probability that at most one breakdown will be observed during the next month is given by the sum of the probabilities of zero and one breakdown. Thus,

$$P(at\ most\ 1\ breakdown) = P(x = 0) + P(x = 1)$$

$$= \frac{3^0 e^{-3}}{0!} + \frac{3^1 e^{-3}}{1!} = \frac{(1)(0.049787)}{1} + \frac{(3)(0.049787)}{1}$$

$$= 0.0498 + 0.1494 = 0.1992.$$

2.2. Continuous Probability Distribution

A random variable that can assume any value contained in one or more intervals is called a continuous random variable.

The following are a few of continuous random variables.

a. Salaries of workers.

b. Time taken by workers to learn a job.
c. Prices of houses.

Example (18)

Classify the following random variables as discrete or continuous.

a. The number of new accounts opened at a bank during a certain month.
b. The time taken by a lawyer to write a real estate contract.
c. The price of concert ticket.
d. The number of workers employed at a randomly selected company.
e. The weight of a randomly selected package.

Solution

a. discrete b. continuous c. continuous d. discrete

2.2.1 Normal Probability Distribution

The normal distribution is one of the many probability distributions that a continuous random variable can possess. The normal distribution is the most important and most widely used off all probability distributions. A large number of phenomena in the real world are normally distributed either exactly or approximately.

The graph of a normal probability distribution with mean μ and standard deviation σ is shown in Figure 3.1 the mean μ locates the center of the distribution, and the distribution is symmetric about its mean μ.

The normal probability distribution has several important properties

1. There is an entire family of normal probability distribution, with the specific shape of each normal distribution being determined by its mean μ and its standard deviation σ.
2. The highest point on the normal curve is located at the mean, which is also the median and the mode of the distribution.
3. The normal distribution is symmetrical, with the curve's shape to the left of the mean being the mirror image of its shape to the right of the mean.

4. Since the normal curve is symmetrical; the area under the normal curve to the right of the mean μ equals the area under the normal curve to the left of the mean.

Science the total area under the normal probability distribution is equal to 1, the symmetry implies that the area to the right of μ is 0.5 and the area to the left of μ is also 0.5. the shape of the distribution is determined by σ, the population standard deviation. Large values of σ reduce the highest of the curve and increase the spread; small values of σ increase the height of the curve and reduce the spread.

2.2.2 The standard Normal Distribution

The standard normal distribution is a special case of the normal distribution.

Standard Normal Distribution

The normal distribution with $\mu = 0$ and $\sigma = 1$ is called the standard normal distribution.

Example (19)

Find the area under the standard normal curve between $z = 0 \, and \, z = 1.95$.

Solution

We divide the number 1.95 into two portions: 1.9 (the digit before the decimal and one digit after the decimal) and 0.05 (the second digit after the decimal). Note that ($1.9 + 0.05 = 1.95$) to find the required area under the standard normal curve, we locate 1.9 in the column for z on the left side of Table (1) and 0.05 in the row for z at the top of Table (1). The entry where the row for 1.9 and the column for 0.05 intersect gives the area under the standard normal curve between z = 0 and z = 1.96, the entry where the row for 1.9 and the column for 0.05 cross is 0.4744.

Example (20)

Find the area to the right of z = 2.32 under the standard normal curve.

Solution

To find the area to the right of z = 2.32, first the area between z $= 0$ and z $= 2.32$ must be found, then subtract this area from 0.5, which is the total area to the right of $z = 0$. From Table (1), the area between $z = 0$ and $z = 2.32$ is 0.4898. Then

$$P(x \geq 2.32) = 0.5 - 0.4898 = 0.0102.$$

Example (21)

Find $P(z \leq -0.5)$. This probability corresponds to the area to the left of a point $z = -0.5$ standard deviation to the left of the mean.

Solution

$$P(z \leq -0.5) = 0.5 - 0.1915 = 0.3085.$$

Example (22)

Find $P(-0.5 \leq z \leq 1.38)$. This probability is the area between $z = -0.5$ and $z = 1.38$.

Solution

$$P(-0.5 \leq z \leq 1.38) = 0.1915 + 0.4162 = 0.6077.$$

As shown from standard normal distribution table, it can be used to find areas under the standard normal curve. However, in real-world applications, a continuous random variable may have a normal distribution with values of the mean and standard deviation different from 0 to 1, the first step, in such a case, is to convert the given normal distribution to the standard normal distribution. This procedure is called ***standardizing a normal distribution***.

The normal random variable x is standardized by expressing its value as the number of standard deviations σ, it lies to the left or right of its mean μ. The standardized normal random variable, z, is defined as

$$z = \frac{x - \mu}{\sigma},$$

where μ and σ are the mean and standard deviation of the normal distribution of x. From the formula for z, these conclusions can be drawn

- when x is less than the mean μ, the value of z is negative.
- when x is greater than the mean μ, the value of z is positive.
- when $x = \mu$ the value of $z = 0$.

Example (23)

Let x be a continuous random variable that has a normal distribution with a mean of 25 and a standard deviation of 4. Find the area

a. $P(25 \leq z \leq 32)$

b. $P(18 \leq z \leq 34)$

Solution

a. $P(25 \leq z \leq 32) = P\left(\frac{25-25}{4} \leq z \leq \frac{32-25}{4}\right)$

$$= P(0 \leq z \leq 1.75) = 0.4599.$$

b. $P(18 \leq z \leq 34) = P\left(\frac{18-25}{4} \leq z \leq \frac{34-25}{4}\right)$

$$= P(-1.75 \leq z \leq 2.25)$$

$$= 0.4599 + 0.4878 = 0.9477.$$

Example (24)

Studies show that gasoline use for compact cars sold in the United States is normally distributed, with a mean of 25.5 miles per gallon (mpg) and a standard deviation of 4.5 mpg. What percentage of compacts get 30 mpg or more?

Solution

$$P(x \geq 30) = P\left(\frac{x-\mu}{\sigma} \geq \frac{30-25.5}{4.5}\right)$$

$$= P(z \geq 1) = 0.5 - 0.3143 = 0.1587$$

Exercises

1. Explain the meaning of the probability distribution of a discrete random variable.
2. Let x denote the number of suits owned by a randomly selected CEO of a corporation. The following table lists the frequency distribution of x for 1000 CEOs

x	4	5	6	7	8	9	10
f	70	180	240	210	170	90	40

   a. Construct a probability distribution table for the number of suits owned by a CEO. Graph the probability distribution.
   b. Find the following probabilities
      i. $P(x = 5)$
      ii. $P(x > 6)$
      iii. $P(4 \leq x \leq 7)$
      iv. $P(x \leq 6)$
3. Let x be the number of errors that a randomly selected page of a book contains. The following table lists the probability distribution of x

x	0	1	2	3	4
$P(x)$	0.73	0.16	0.06	0.04	0.01

Find the mean and standard deviation of x

4. Let x be a discrete random variable that possesses a binomial distribution. Using the binomial formula, find the following probabilities.
   a. $P(x = 5)$ for $n = 8$ and $p = 0.60$
   b. $P(x = 3)$ for $n = 4$ and $p = 0.30$
   c. $P(x = 2)$ for $n = 6$ and $p = 0.20$
5. Thirty percent of all customers who enter a store will make a purchase. Suppose that six customers enter the store and that these customers make independent purchase decisions.
   a. Let $x =$the number of the six customers who will make a purchase. Write the binomial formula for this situation.
   b. Use the binomial formula to calculate:
      (1) The probability that exactly five customers make a purchase.
      (2) The probability that at least three customers make a purchase.
      (3) The probability that two or fewer customers make a purchase.

(4) The probability that at least one customer makes a purchase

6. Explain the hyper-geometric probability distribution. Under what conditions is this probability distribution applied to find the probability of a discrete random variable x?
7. Let $N = 14, r = 6, and\ n = 5.$ Using the hypergeometic probability distribution formula, find

   a. $P(x = 5)$ b. $P(x = 0)$ c. $P(x \leq 1)$
8. What are the conditions that must be satisfied to apply the Poisson probability distribution?
9. What is the parameter of the Poisson probability distribution, and what does it mean?
10. A commuter airline receives an average of 9.7 complaints per day from its passengers. Using the Poisson formula, find the probability that on a certain day this airline will receive exactly 7 complaints.
11. Let x be a Poisson random variable with mean $\mu = 2$. Calculate these probabilities:

    a. $P(x = 0)$ b. $P(x = 1)$

    c. $P(x > 1)$ d. $P(x = 5)$

12. Calculate the area under the standard normal curve to the left of these values

a. $z = 1.6$ b. $z = 1.83$

c. $z = 0.90$ d. $z = 4.18$

13. Find the following probabilities for the standard normal random variable

a. $P(-1.43 < z < 0.68)$

b. $P(0.58 < z < 1.74)$

c. $P(-1.55 < z < -0.44)$

d. $P(z > 1.34)$

e. $P(z < -4.32)$

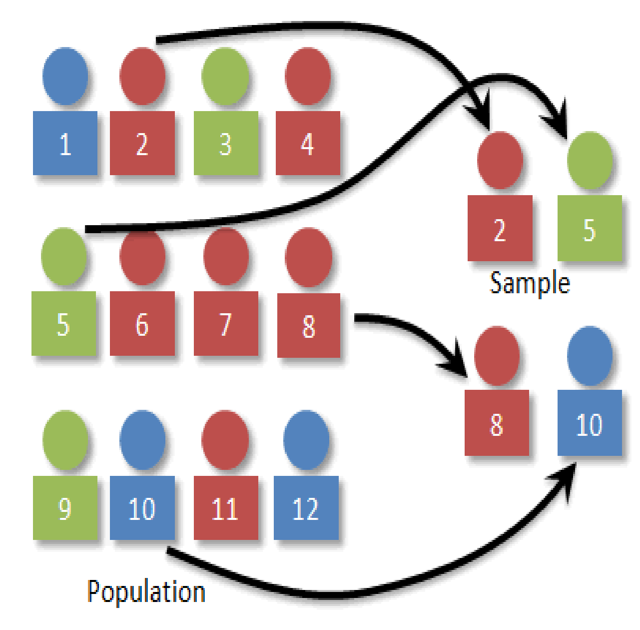
14. If the student's weight in the military academy is 65 k.g. with a standard deviation 5 k.g. suppose we selected a student randomly, find the following probabilities.

a. The student weight falls between 65 and 70 k.g.

b. The student weight falls between 65 and 85 k.g.

c. The student weight falls between 60 and 70 k.g.

d. The student weight is less than 75 k.g.



CHAPTER

Three

Sampling Distributions

3.1 Sampling Plans

3.2 The sampling distribution of the sample mean

3.3 Mean and Standard Deviation of $\bar{X}$

3.4 Mean and Standard Deviation of Sample Proportions

Sampling Distributions

In this chapter, the properties of two important sampling distributions, the sampling distribution of the sample mean and the sampling distribution of the sample proportion will be discussed.

Usually, to conduct research, we select a portion of the target population is called a sample. Then we collect the required information from the elements included in the sample.

3.1 Sampling Plans

The way a sample is selected is called the ***sampling plan or experimental design***. The technique of collecting information from a portion of the population is called a ***sample survey***.

The way a sample is selected is called the sampling plan or experimental design and determines the quantity of information in the sample. Depending on how a sample is drawn, it may be a ***random sample*** or ***nonrandom sample.***

Random and Nonrandom samples

A random sample is a sample drawn in such a way that each member of the population has some chance for being selected in the sample. In a nonrandom sample, some members of the population may not have any chance of being selected in the sample.

Simple random sample

If a sample of n elements is selected from a population of N elements using a sampling plan in which each of the possible samples has the same chance of selection, then the sampling is said to be random and the resulting sample is a ***simple random sample***.

In addition to simple random sampling, there are other sampling plans that involve randomization and therefore provide a probabilistic basis for inference making.

Stratified random sample

Stratified random sampling involves selecting a simple random sample from each of a given number of subpopulations, or **strata**.

Another form of random sampling is used when the available sampling units are groups of elements, called ***clusters***.

A Cluster Sample

A cluster sample is a simple random sample of clusters from the available clusters in the population.

Sometimes the pollution to be sampled is ordered, such as an alphabetized list of people with driver's licenses, one element is chosen at random from the first K element, and then every kth element therefore is included in the sample.

A 1 in k Systematic Random Sample

A 1 in k Systematic Random Sample involves the random selection of one of the first k element in an ordered population, and then the systematic selection of every kth element thereafter.

3.2 The sampling distribution of the sample mean

The sampling distribution of the sample mean $\bar{X}$ is the probability distribution of the population of all possible means obtained from samples of size n.

Example (1)

A population consists of the following four values: 2, 2, 4 and 6:

a. List all samples of size 2, and compute the mea of each sample.
b. Compute the mean of the distribution of sample means and the population mean compare the two values.

c. Compare the distribution in the population with that of the sample means.

Solution

X: 2, 2, 4, and 6

Sample	Values	Mean
1	2,2	2
2	2,4	3
3	2,6	4
4	2,4	3
5	2,6	4
6	4,6	5

$$\mu = \frac{2 + 2 + 4 + 6}{4} = 3.5$$

$$\bar{\bar{X}} = \frac{\sum \bar{X}}{No\ of\ samples} = \frac{21}{6} = 3.5$$

Example (2)

A population consists of N=5 numbers: 3, 6, 9, 12, 15. If a random sample of size $n = 3$ is selected without replacement,

a. List all samples of size 3, and compute the mean of each sample.
b. Compute the mean of the distribution of sample means and the population mean compare the two values.
c. Compare the distribution in the population with that of the sample means.

Solution

a.

Sample	Values	Mean
1	3,6,9	6
2	3,6,12	7
3	3,6,15	8
4	3,9,12	8
5	3,9,15	9

6	3,12,15	10
7	6,9,12	9
8	6,9,15	10
9	6,12,15	11
10	9,12,15	12

$$b.\ \mu = \frac{3+6+9+12+15}{5} = 9.$$

$$\bar{\bar{X}} = \frac{\sum \bar{X}}{No\ of\ samples} = \frac{90}{10} = 9$$

c. The dispersion of the population is greater than that of the same means. The sample means vary from 6 to 12, whereas the population varies from 3 to 15.

3.3 Mean and Standard Deviation of $\bar{X}$

The mean and standard deviation of the sampling distribution of $\bar{x}$ are called the mean and standard deviation of $\bar{x}$ and are denoted by $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$, respectively.

Mean of the Sampling distribution of $\bar{X}$

The mean of the sampling distribution of $\bar{x}$ is equal to the mean of the population. Thus

$$\mu_{\bar{x}} = \mu,$$

Standard Deviation of the Sampling distribution of $\bar{X}$

The standard deviation of the sampling distribution of $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population and n is the sample size.

The above formula holds true if the sample size is small in comparison to the population size. The sample size is considered to be small compared to the population size if the sample size is equal to or less than 5% of the population size, that is, if

$$\frac{n}{N} \leq 0.05.$$

If the condition is not satisfied, the following formula can be used to calculate $\sigma_{\bar{x}}$.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

where the factor $\sqrt{\frac{N-n}{N-1}}$ is called the finite population correction factor.

Example (3)

The mean wage per hour for all 5000 employees working at a large company is \$13.50 with a standard deviation of \$2.90. Let $\bar{x}$ be the mean wage per hour for a random sample of certain employees selected from this company. Find the mean and standard deviation of $\bar{x}$ for a sample size of

(a) 30 (b) 75 (c) 200

Solution

(a) $\mu_{\bar{x}} = \mu = \$13.50,$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.90}{\sqrt{30}} = \$0.53.$$

(b) $\mu_{\bar{x}} = \mu = \$13.50,$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.90}{\sqrt{75}} = \$0.33.$$

(b) $\mu_{\bar{x}} = \mu = \$13.50,$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.90}{\sqrt{200}} = \$0.21.$$

From the above calculations, the mean of the sampling distribution of $\bar{x}$ is always equal to the mean of the population whatever the size of the sample. However, the value of the standard deviation of $\bar{x}$ decrease from \$ 0.53 to \$0.33 and then to 0.21 as the sample size increases from 30 to 75 and then to 200.

3.4 Mean and Standard Deviation of Sample Proportions

The population proportion, denoted by P, is obtained by taking the ratio of the number of elements in a population with a specific characteristic to the total number of elements in the population.

Population and Sample Proportions

The population and sample proportions, denoted by p and $\hat{p}$, respectively, are calculated as

$$P = \frac{x}{n} \quad \text{and} \quad \hat{p} = \frac{x}{n}$$

where

$N =$ Total number of elements in the population.

$n =$ Total number of elements in the sample.

$x =$ Number of elements in the population or sample that possess a specific characteristic.

Mean and Standard Deviation of $\hat{p}$

The mean of $\hat{p}$, which is the same as the mean of the sampling distribution of $\hat{p}$, is always equal to the population proportion P just as the mean of the sampling distribution of $\bar{x}$ is always equal to the population mean μ.

Mean of the Sample Proportion

The mean of the sample proportion $\hat{p}$ is obtained by $\mu_{\hat{p}}$ and is equal to the population proportion P. Thus,

$$\mu_{\hat{p}} = p.$$

Standard Deviation of the Sample Proportion

The standard deviation of the sample of the sample proportion $\hat{p}$ is denoted by $\sigma_{\hat{p}}$ and is given by the formula

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

where P is the population proportion, $q = 1 - p$, and n is the sample size. This formula is used when ${}^{n}/_{N} \leq 0.05$ where N is the population size.

Example (4)

Suppose a total of 789,654 families live in a city and 563,282 of them own homes. Then,

$N =$ population size = 789,654

$X =$ families in the population who own homes$=$ $563{,}282$

The proportion of all families in this city who own homes is

$$p = \frac{x}{N} = \frac{563{,}282}{789{,}654} = 0.71$$

$n =$ population size =240

$x =$ Families in the population who own homes= 158

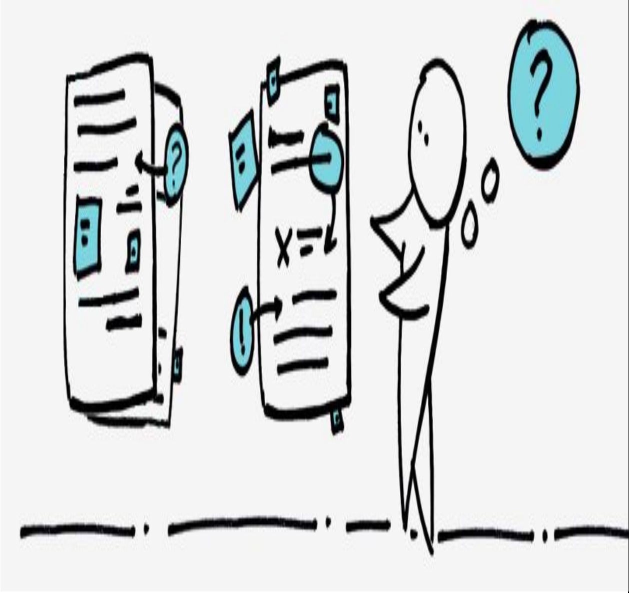
$$\hat{p} = \frac{x}{n} = \frac{158}{240} = 0.66$$

Exercises

1. Let $\bar{x}$ be the mean of a sample selected from a population
   a. What is the mean of $\bar{x}$ equal to?
   b. What is the standard deviation of $\bar{x}$ equal to?
2. Consider a large population with $\mu = 90$ and $\sigma = 16$. Assuming $n/N \leq 0.05$, find the mean and standard deviation of the sample mean for a sample size of
   a. 18 b. 90
3. A population of $N = 5000$ has $\sigma = 20$. In each of the following cases which formula will you use to calculate $\sigma_{\bar{x}}$ and why? Using the appropriate formula, calculate $\sigma_{\bar{x}}$
   a. $n = 500$ b. $n = 200$ c. $n = 500$ d. n= 100
4. A population consists of N=5 numbers: 10, 10, 12, 13, 15. If a random sample of size $n = 3$ is selected without replacement,
   a. List all samples of size 3, and compute the mean of each sample.

b. Compute the mean of the distribution of sample means and the population mean compare the two values.
c. Compare the distribution in the population with that of the sample means.

5. The amounts of electric bills for all households in a city have a skewed probability distribution with a mean of \$65 and a standard deviation of \$25. Find the probability that the mean amount of electric bills for a random sample of 75 households selected from this city will be
   a. More than \$70
   b. Between \$58 and 63
   c. More than the population mean by at least \$5



CHAPTER	
Four	

Estimation of the mean and proportion

- **4.1. Types of Estimators**
- **4.2. Interval Estimation of A Population Mean: (Large Samples)**
- **4.3. Interval Estimation of A Population Proportion: (Large Samples)**
- **4.4. Tests of Hypothesis**
- **4.5. Analysis of variance**

Estimation of the mean and proportion

Inferential statistics is defined as a part of statistics that helps us to make decisions about some characteristics of a population based on sample information. Methods for making inferences about population parameters fall into one of two categories:

- Estimation: Estimating or predicting the value of the parameter.
- Hypothesis testing: Making a decision about the value of a parameter based on some preconceived idea about what its value might be.

4.1. Types of Estimators

To estimate the value of a population parameters, information from the sample in the form of an estimator can be used.

Estimate and Estimator

The value assigned to a population parameter based on the value of a sample statistics is called an estimate. The sample statistics used to estimate a population parameter is called an estimator.

Estimators are used in two different ways:

- Point Estimator: the value of a sample statistics that is used to estimate a population parameter is called a point estimator.
- Interval Estimator: Based on sample data, two numbers are calculated to form an interval within which the parameter is expected to lie. The formula that describes this calculation is called the interval estimator.

Unbiased Estimator

An estimator of a parameter is said to be unbiased if the mean of its distribution is equal to the true value of the parameter. Otherwise, the estimator is said to be biased.

Error of Estimation

The distance between an estimate and estimated parameter is called the error of estimation.

Usually, whenever the point estimation is used, the margin of error associated with that point estimation is calculated. For the estimation of the population mean, the margin of error is calculated as follows.

$$\text{Margin of error} = \mp 1.96\, \sigma_{\bar{X}} \quad \text{or error} = \mp 1.96\, s_{\bar{X}}$$

4.2. Interval Estimation of A Population Mean: (Large Samples)

Each interval is constructed with regard to a given confidence level and is called a confidence interval. The confidence level associated with a confidence interval states how much confidence we have that interval contains the true population parameter. To construct a confidence interval for the population mean μ when the sample size is large. When the sample size is 30 or large, the normal distribution to construct a confidence interval for μ is used. The confidence interval for μ is

$$\bar{x} \mp z\sigma_{\bar{x}} \qquad \text{if } \sigma \text{ is known}$$

$$\bar{x} \mp zs_{\bar{x}} \qquad \text{if } \sigma \text{ is not known}$$

where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ and $s_{\bar{x}} = s/\sqrt{n}$.

The value z is read from the standard normal distribution table for the given confidence level.

Example (1)

A scientist interested in monitoring chemical contaminants in food, and thereby the accumulation of contaminants in human diets, selected a random sample of n = 50 male adults. It was found that the average daily intake of dairy products was $\bar{x} = 756$ grams per day with a standard deviation of $s = 35$ grams per day. Use this sample information to construct

a. 95% confidence interval for the mean daily intake of dairy products for men.
b. 99% confidence interval for the mean daily intake of dairy products for adult men.

Solution

a. $$\bar{x} \mp z s_{\bar{x}} = \bar{x} \mp z\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} + z\left(\frac{s}{\sqrt{n}}\right) = 756 + 1.96\left(\frac{35}{\sqrt{50}}\right) = 765.70$$

$$\bar{x} - z\left(\frac{s}{\sqrt{n}}\right) = 756 - 1.96\left(\frac{35}{\sqrt{50}}\right) = 746.30$$

Hence, the 95% confidence interval for μ is from 746.30 to 765.70 grams per day.

b. $$\bar{x} + z\left(\frac{s}{\sqrt{n}}\right) = 756 + 2.58\left(\frac{35}{\sqrt{50}}\right) = 768.77$$

$$\bar{x} - z\left(\frac{s}{\sqrt{n}}\right) = 756 - 2.58\left(\frac{35}{\sqrt{50}}\right) = 743.23$$

The 99% confidence interval for μ is from 743.23 to 768.77 grams per day.

4.3. Interval Estimation of A Population Proportion: (Large Samples)

Many research experiments or sample surveys have as their objective the estimation of the proportion of a people or objects in a large group that possess a certain characteristic. For example, the production manager of a company may want to estimate the proportion of defective items produced on a machine.

The population proportion is denoted by p and the sample proportion is denoted $\hat{p}$

- The sampling distribution of the sample proportion $\hat{p}$ is (approximately) normal.
- The mean $\mu_{\hat{p}}$ of the sampling distribution of $\hat{p}$ is equal to the population proportion p.
- The standard deviation $\sigma_{\hat{p}}$ of the sampling distribution of the sample proportion $\hat{p}$ is $\sqrt{pq/n}$ where $q = 1 - p$.

The value of $s_{\hat{p}}$, which is a point estimate of $\sigma_{\hat{p}}$, is calculated as

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

The sample proportion $\hat{p}$ is the point estimator of the corresponding population proportion, the margin of error associated with that point estimation is calculated as follows:

$$\text{Margin of error} = \mp 1.96 s_{\hat{p}}$$

The confidence interval for the population proportion p is

$$\hat{p} \mp z s_{\hat{p}}$$

The value of z used here is obtained from the standard normal distribution table for the given confidence level.

where $\hat{p} = \frac{x}{n} = \frac{Total\, number\ of\ successes}{Total\, number\ of\ trails}$, and the sample size is considered large when the normal approximation to the binomial distribution is adequate-namely, when $n\hat{p} > 5$ and $n\hat{q} < 5$.

Example (2)

A random sample of 985 "likely" voters-those who are likely to vote in the upcoming election were polled during a phone-athon conducted by the Republican Party. Of those surveyed, 592 indicated that they intended to vote for the Republican candidate in the upcoming election. Construct a 90% confidence interval for p, the proportion of likely voters in the population who intended to vote the Republican candidate. Based on this information, can you conclude that the candidate will win the election?

Solution

The point estimate for p is

$$\hat{p} = \frac{x}{n} = \frac{592}{985} = 0.601$$

and the standard error is

$$\sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.601)(0.399)}{985}} = 0.016$$

$$\hat{p} + zs_{\hat{p}} = 0.601 + 0.026 = 0.627$$

$$\hat{p} - zs_{\hat{p}} = 0.601 - 0.026 = 0.575$$

or $0.575 < p < 0.627$. the percentage of likely voters who intend to vote for the Republican candidate is between 57.5% and 62.5%.

4.4 Tests of Hypothesis

A Hypothesis is a statement about the value of a population parameter developed for the purpose of testing. Example of hypotheses made about a population parameter is the mean monthly income for systems analysts is 3625.

Null Hypothesis

A null hypothesis is a claim (or statement) about a population parameter that is assumed to be true until it is declared false.

Alternative Hypothesis

An alternative hypothesis is a claim about a population parameter that will be true if the null hypothesis is false.

		Actual Situation	
		H_0 is true	H_0 is false
Decision	Don't reject H_0	Correct Decision	Type II or β error
	Reject H_0	Type I or $\propto$ error	Correct Decision

Type I Error

A Type I error occurs when a true null hypothesis is rejected. The value of α represents the probability of committing this type of error. The value of α represents the significance level of the test.

Type II Error

A Type II error occurs when a false null hypothesis is not rejected. The value of β represents the probability of committing a type II error.

	Two-tailed test	Left-tailed test	Right-tailed test
Sign in the null hypothesis	$=$	$= or\ \geq$	$= or\ \leq$
Sign in the alternative hypothesis	$\neq$	$<$	$>$
Rejection region	In both tails	In the left tail	In the right tail

Steps of Testing of Hypothesis

Step 1: State the null hypothesis $\boldsymbol{H_0}$ and the alternate hypothesis $\boldsymbol{H_1}$.

Step 2: Select the level of significance, α = 0.01 as stated in the problem

Step 3: Select the test statistic.

Use Z-distribution

If the population standard deviation is known or the sample is greater than 30.

Use *t*-distribution

If the population standard deviation is unknown and the sample is less than 30.

Step 4: Formulate the decision rule.

$$\text{Reject H0 if } |Z| > Z\alpha/2$$

or graph and determine acceptance area and reject area then take the decision

Example (3)

The mean of a normally distributed population is believed to be equal to 50.1. A sample of 36 observations is taken, and the sample mean is found to be 53.2. the alternative hypothesis is that the population mean is not equal to 50.1. Complete the hypothesis test, assuming that the population standard deviation is equal to 4. Use a 0.05 significance level. (Z=1.96)

Solution

Step 1: Determine the null hypothesis and alternative hypothesis

$$H_0: \mu = 50.1$$

$$H_1: \mu \neq 50.1$$

Step 2: Select the level of significance

$$Z_{tab} = 1.96, \quad \alpha = 0.05$$

Step 3: Identify the test statistic

$$Z_{cal} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{53.2 - 50.1}{4/6} = 4.65$$

Step 4: Formulate a decision rule

Reject H_0 if $Z_{cal} > Z_{tab}$

$4.65 > 1.96$ Reject H_0 and accept H_1.

Example (4)

The manufacturer of the X-15 steel-belted radial truck tire claims that the mean mileage the tire can be driven before the thread wears out is at least 80,000 km. The standard deviation of the mileage is 8,000 km. The Crosset Truck Co. bought 48 tires and found that the mean mileage for their trucks is 79,200 km. From Crosset's experience, can we conclude, with 5% significance level, that the manufacturer's claim is wrong? ($Z_{tab} = 1.65$)

Solution

Step 1: Determine the null hypothesis and alternative hypothesis

$$H_0: \mu \geq 80{,}000$$

$$H_1: \mu < 80{,}000$$

Step 2: Select the level of significance

$$Z_{tab} = 1.65,$$

Step 3: Identify the test statistic

$$Z_{cal} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{79{,}200 - 80{,}000}{8{,}000/\sqrt{48}} = -0.693$$

Step 4: Formulate a decision rule

Reject H_0 if $|Z_{cal}| < Z_{tab}$

$0.693 < 1.65$ Don't reject H_0 and reject H_1.

Example (5)

Consider the following null and alternative hypotheses:

$$H_0: \mu = 90$$

$$H_1: \mu \neq 90$$

A random sample of 20 observations taken from this population produced a **sample** **mean of 85**. The population **standard deviation** is known to be **7**. a. If this test is made at the 1% significance level, would you reject the null hypothesis? $\alpha = 0.01(\boldsymbol{z = 2.58})$?

Solution

State the null and alternative hypotheses.

$$H_0: \mu = 90\, minuts$$

$$H_1: \mu \neq 90\, minuts$$

Use Z-distribution since σ is known

$$z_{cal} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$z_{cal} = \frac{85 - 90}{\frac{7}{\sqrt{20}}} = -3.19$$

<u>Formulate the decision rule</u>

$$|z_{cal}| < z_{tab}$$

$$3.19 \nless 2.58$$

Reject H_0 and accept H_1

4.5 One Way Analysis of Variance

The analysis of variance procedure is used to test the null hypothesis that means of three or more populations are the same against the alternative hypothesis that all population means are not equal.

<u>ANOVA</u> is a procedure used to test the null hypothesis that the means of three or more populations are equal.

The **<u>ANOVA</u>** test is applied by calculating two estimates of the variance σ^2 of population distributions

- **The variance between samples.**
- **The variance within samples.**

The variance between samples is called the mean square between samples or MSB. The variance within samples is called the mean square within samples or MSW.

<u>Test Statistic F for A One Way of ANOVA</u>
The value of the test statistic F for an ANOVA test is calculated as

$$\boldsymbol{F = \frac{Variance\ between\ samples}{Variance\ within\ samples} = \frac{MSB}{MSW}}$$

Example (6)

One of the products made at Abe Chemicals Company is detergents. Due to increased sales, the company is planning to buy a few new machines that will be used to fill the 64-ounces detergent jugs, the company is considering three types of such machines but will eventually buy only type of machine. Before making such a decision, the company wanted to test the three machines for the number of jugs filled per hour. To do so, the company used each of the three types of machines for five hours and recorded the number of jugs filled during each hour by these machines. The following table gives the number of jugs filled by these machines during each of the five hours

Machine 1	Machine 2	Machine 3
54	53	49
49	56	53
52	57	47
55	51	50
48	59	54

$\mathbf{x} = $ **the number of jugs filled by a machine during a given hour.**

$\mathbf{k} = $ **the number of different machines (or treatment).**

$\mathbf{n_i} = $ **the sample size of sample i. (The number of hours for which each machine is used).**

$\mathbf{T_i} = $ **the sum of the values in sample i.**

$\mathbf{n} = $ **the number of values in all samples**

$$\mathbf{n = n_1 + n_2 + n_3 + \cdots}$$

$\sum \mathbf{x} = $ **the sum of the values in all samples**

$$\sum \mathbf{x = T_1 + T_2 + T_3 + \cdots}$$

$\sum \mathbf{x^2} = $ **the number of the squares of the values in all samples.**

To calculate MSB and MSW, we first compute the between-samples sum of squares denoted by SSB and within-samples sum of squares denoted by SSW. The sum of SSB and SSW is called the total sum of squares and it is denoted by SST, that is,

$$SST = SSB + SSW$$

Between and within samples sum of squares

The between samples sum of squares, denoted by SSB, is calculated as

$$SSB = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \cdots\right) - \frac{(\sum x)^2}{n}$$

The within-sample sum of squares, denoted by SSW, is calculated as

$$SSW = \sum x^2 - \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \cdots\right)$$

Machine 1	Machine 2	Machine 3
54	53	49
49	56	53
52	57	47
55	51	50
48	59	54
$T_1 = 258$	$T_2 = 276$	$T_3 = 253$
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$

$$\sum x = \mathrm{T}_1 + \mathrm{T}_2 + \mathrm{T}_3 = 258 + 276 + 253 = 787$$

$$n = \mathrm{n}_1 + \mathrm{n}_2 + \mathrm{n}_3 = 5 + 5 + 5 = 15$$

$$\sum x^2 = (54)^2 + (49)^2 + (52)^2 + (55)^2 + (48)^2 + (53)^2 + (56)^2 + (57)^2 + (51)^2 + (59)^2 (49)^2 + (53)^2 + (47)^2 + (50)^2 + (54)^2 = 41,461.$$

Substituting all the values in the formula for SSB and SSW, we obtain the following values of SSB and SSW

$$SSB = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3}\right) - \frac{(\sum x)^2}{n}$$

$$SSB = \left(\frac{(258)^2}{5} + \frac{(276)^2}{5} + \frac{(253)^2}{5}\right) - \frac{(787)^2}{15} = 58.5333$$

$$SSW = \sum x^2 - \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3}\right)$$

$$SSW = 41,461 - \left(\frac{(258)^2}{5} + \frac{(276)^2}{5} + \frac{(253)^2}{5}\right) = 111.2000$$

$$SST = 58.5333 + 111.2000 = 169.7333$$

The variance between samples MSB and the variance within samples MSW are calculated using the following formulas.

MSB and MSW

$$MSB = \frac{SSB}{k-1},$$

$$MSW = \frac{SSW}{n-k},$$

and

$$F = \frac{MSB}{MSW}$$

$$MSB = \frac{SSB}{k-1} = \frac{85.5333}{3-1} = 29.2667$$

$$MSW = \frac{SSW}{n-k} = \frac{111.2000}{15-3} = 9.2667$$

$$F = \frac{MSB}{MSW} = \frac{29.2667}{9.2667} = 3.16$$

ANOVA Table

Source of variation	Sum of squares	Degree of freedom	Mean square	F
Between	**SSB**	**k-1**	**MSB**	$F = \frac{MSB}{MSW}$
within	**SSW**	**n-k**	**MSW**	
Total	**SST**	**n-1**		

Source of variation	Sum of squares	Degree of freedom	Mean square	F
Between	**2**	**58.5333**	**29.2667**	**3.16**
within	**12**	**111.2000**	**9.2667**	
Total	**14**	**169.7333**		

To make a test about the equality of means of three populations, we follow our standard five-step procedure

Step 1: State the null hypothesis and the alternative hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$(the\ means\ of\ the\ three\ machines\ are\ equal).$$

$$H_1: all\ three\ means\ are\ not\ equal$$

Step 2: Select the distribution to use

Because we are comparing three means for three normally distributed populations, we use the F distribution to make the test

Step 3: Determine the rejection and non-rejection regions.

At the 0.01 significance level, is there a difference in the mean scores? where $F_{(0.01,2,12)} = 6.93$.

Step 4: Calculate the values of the test statistics

$$F = 3.16$$

Step 5: Make a decision

$$F > F_{(0.01,2,12)}???$$

Because the value of the test statistic $F = 3.16$ is less than the critical value of $F_{(0.01,2,12)} = 6.93$, it falls in the non-rejection region.

We fail to reject H_0, and conclude that the three means are equal.

Exercises

1. Explain the difference between an estimator and an estimate.
2. Explain the meaning of a point estimate and an interval estimate.
3. What is the point estimator of the population mean μ.
4. Explain what is meant by margin of error in point estimation.
5. Calculate the margin of error in estimating a population mean μ for these values
   a. $n = 30, \sigma^2 = 0.2$
   b. $n = 30, \sigma^2 = 0.9$
   c. $n = 30, \sigma^2 = 1.5$
   d. What effect does a larger population variance have on the margin of error?
6. Calculate the margin of error in estimating a binomial proportion for each of the following values of n. Use $p = 0.5$ to calculate the standard error of the estimator.

a. $n = 30$
b. $n = 100$
c. $n = 400$
d. What effect does increase the sample size have on the margin of error?

7. The standard deviation for a population a population is $\sigma = 12.6$. A sample of 36 observations selected from this population gave a mean equal to 74.8.
   a. Make a 90% confidence interval for μ
   b. Construct a 95% confidence interval for μ
   c. Determine a 99% confidence interval for μ
   d. Does the width of the confidence intervals construct in a through c increase as the confidence level increases? Explain your answer.
8. For a sample data set, $\bar{x} = 16\ and\ s = 5.3$
   a. Construct a 95% confidence interval for μ assuming $n = 50$
   b. Construct a 90% confidence interval for μ assuming $n = 50$.

CHAPTER

Five

Regression and Correlation

5.1 Regression

5.2. Correlation

Regression and Correlation

A regression model is a mathematical equation that describes the relationship between two or more variables. A simple regression model includes only two variables: one independent and one dependent. The dependent variable is the one being explained and the independent variable is the one used to explain the variation in the dependent variable.

5.1 Regression

The relationship between two variables in a regression is expressed by a mathematical equation called a regression model or regression equation.

In a regression model, the independent variable is usually denoted by x and the dependent variable is usually denoted by y.

Equation of A regression Model

In the regression model $y = A + Bx + \varepsilon, A$ is called the y-intercept or constant term, B is the slope, and ε is the error term. The dependent and independent variables are y and x, respectively. A and B are the population parameters. The regression line obtained for the regression model is called population regression line. The values of A and B in the population regression line are called the true values of the y-intercept and slope.

However, population data are difficult to obtain. As a result, sample data is used to estimate the regression model. The values of the y-intercept and the slope calculated from sample data on x and y, the estimated model is given by

$$\hat{y} = a + bx$$

Estimate of A and B

In the model $\hat{y} = a + bx$, a and b, which are calculated using sample data, are called the estimate of A and B.

$$a = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2},$$

and

$$b = \bar{y} - a\bar{x}$$

Example 1

The numbers (in millions) between consumption y and production x are recorded as follows:

x	3	6	5	4	5	2	10
y	3	4	2	6	5	1	7

Determine the regression equation and predict the consumption value when the production $= 20$.

Solution

x	y	x^2	y^2	xy
3	3	9	9	9
6	4	36	16	24
5	2	25	4	10
4	6	16	36	24
5	5	25	25	25
2	1	4	1	2
10	7	100	49	70
35	**28**	**215**	**140**	**164**

$$\hat{Y} = ax + b$$

$$\bar{x} = \frac{35}{7} = 5, \qquad \bar{y} = \frac{28}{7} = 7$$

$$a = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} = \frac{164 - 5 \times 5 \times 7}{215 - 5(25)} = \frac{-11}{90} = -0.12$$

$$b = \bar{y} - a\bar{x} = 7 + 0.12(5) = 7.6$$

$$\hat{Y} = -0.12x + 7.6$$

$$\hat{Y} = -0.12(20) + 7.6 = 5.2$$

5.2 Linear Correlation

Another measure of the relationship between two variables is the correlation coefficient which measures the strength of the linear association between two variables. In other words, the linear correlation coefficient measures how closely the points in a scatter diagram are spread around the regression line. The value of the correlation coefficient always lies between -1 and 1, that is , $-1 \leq \rho \leq +1,$ and $-1 \leq r \leq +1,$

where ρ is the population correlation coefficient and r is the sample correlation coefficient.

If $r = 1$, it is said to be a case of perfect positive linear correlation. In such a case, all points in the scatter diagram lie on a straight line that slopes from the left to right. If $r = -1$, the correlation coefficient is said to be a perfect negative linear correlation. In this case, all points in the scatter diagram fall on a straight line that slopes downward from left to right. If $r = 0$, there is no correlation between variables.

If the correlation between two variables is positive and close to 1, we say that the variables have a strong positive linear correlation. If the correlation between two variables is positive and close to zero, then the variables have a weak positive linear correlation. On the other hand, if the correlation between two variables is negative and close to -1, then the variables are said to have a strong negative linear correlation. Also, if the correlation between variables is negative but close to zero, there is a weak negative linear correlation.

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{[(\sum x^2) - n(\bar{x})^2][(\sum y^2) - n(\bar{y})^2]}}$$

Example 2

Determine the coefficient of correlation between x, y and decide the correlation type for the following data:

X	**3**	**4**	**4**	**5**	**8**
Y	**2**	**3**	**1**	**6**	**7**

Solution

x	y	x^2	y^2	xy
3	2	9	4	6
4	3	16	9	12
4	1	16	1	4
5	6	25	36	30
8	7	64	49	56
24	**19**	**130**	**99**	**108**

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{[(\sum x^2) - n(\bar{x})^2][(\sum y^2) - n(\bar{y})^2]}}$$

$$\bar{x} = \frac{24}{5} = 4.8$$

$$\bar{y} = \frac{19}{5} = 3.8$$

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{[(\sum x^2) - n(\bar{x})^2][(\sum y^2) - n(\bar{y})^2]}} =$$

$$r = \frac{108 - 5 \times 4.8 \times 3.8}{\sqrt{[130 - 5(4.8)^2][99 - 5(3.8)^2]}} = \frac{16.8}{\sqrt{14.8 \times 26.8}}$$
$$= 0.84$$

Example 3

- The number (in millions) between consumption y and production x are recorded as follows

X	35	49	21	39	15	28	25
Y	9	15	7	11	5	8	9

1. Find the correlation coefficient between x and y.
2. Decide the correlation type.
3. Determine the regression equation.
4. Predict the consumption value when x = 15.

X	Y	XY	X^2	Y^2
35	9	315	1225	81
49	15	735	2401	225
21	7	147	441	49
39	11	429	1521	121
15	5	75	225	25
28	8	224	784	64
25	9	225	625	81
$\sum X$ =212	$\sum Y =$ 64	$\sum XY =$ 2150	$\sum X^2 =$ 7222	$\sum Y^2$ =646

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

$$r = \frac{7(2150) - (212)(64)}{\sqrt{[7(7222) - 212^2][7(646) - 64^2]}}$$

$$r = \frac{15050 - 13568}{\sqrt{[50554 - 44944][4522 - 4096]}}$$

$$= \frac{1482}{\sqrt{(5610)(426)}}$$

$$r = \frac{1482}{1545.9} = 0.95, \quad \boldsymbol{strong}, \boldsymbol{direct}$$

$$\boldsymbol{a} = \frac{n\sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$\boldsymbol{a} = \frac{7(2150) - (212)(64)}{[7(7222) - 212^2]} = \frac{1482}{5610} = 0.26$$

$$\bar{x} = \frac{212}{7} = 30.29$$

$$\bar{y} = \frac{64}{7} = 9.14$$

$$b = 9.14 - 0.26(30.29) = 1.26$$

$$\hat{Y} = 0.26x + 1.26$$

$$\hat{Y} = 0.26(15) + 1.26 = 391.26$$

Supplementary Exercises

Question (1): Choose the correct answer

- **A card is drawn at random from a deck of cards, the probability of getting a spade is ………………**

A. 4/52

B. 1/52

C. 13/52

Answer C

If you roll two dice, then the probability that

- **the sum = 12 is**

A. 1/36

B. 1/9

C. 1/18

Answer A

- **The sum = 12 or 11 is**

A. 1/12

B. 1/36

C. 2/52

Answer A

- **Both of the two dice are the same number**

  A. 3/36

  B. 6/36

  C. 2/36

Answer B

- **The sum = 8, given that the first die shows an odd number**

  A. 3/36

  B. 1/18

  C. 1/9

Answer C

- **The following table explains the probability distribution for (x)**

X	0	1	3	4
P(X)	0.3	0.5	0.1	0.1

- **The mean is ……………..**

  A. 1.2

  B. 1.44

  C. 2

Answer A

- **The variance is ……………….**

  A. 7.5

  B. 1.56

  C. 6.06

Answer B

Question (2): Decide if each of the following statement is True or False

- **The probability of an event is_always in the range zero to 1.0.**

  ❑ True

  ❑ False

Answer True

- **Two equally likely events have no effect of each other.**

  ❑ True

  ❑ False

Answer False

- **Two mutually exclusive events cannot occur together.**

  ❑ True

  ❑ False

Answer True

- **A final outcome of an experiment is called a compound event.**

  ❑ True

  ❑ False

Answer False

- **The normal distribution is a discrete probability distribution.**

  ❑ True

  ❑ False

Answer False

- **Sample space is a collection of all outcomes of an experiment.**

  ❑ True

  ❑ False

Answer True

- **The Probability of an event is a number that measures the chance that the event will occur.**

  ❑ True

  ❑ False

Answer True

- **If A and B are two events, then the conditional probability of A given B is written as $\boldsymbol{P(A \cap B)}$.**

  ❑ True

  ❑ False

Answer False

Question (3) Choose the Correct Answer

1. **Descriptive statistics is a collection of observations or measurements on a variable.**

   A. True

   B. False

ANSWER: B

2. **Data or data set is a collection of methods for organizing, displaying, and describing data using tables, graphs, and summary measures.**

A. True

B. False

ANSWER: B

3. **Discrete variable is a quantitative variable whose values are countable.**

A. True

B. False

ANSWER: A

4. **Element or member is a specific subject or object included in a sample or population.**

A. True

B. False

ANSWER: A

5. **Inferential statistics is a collection of methods that help make decisions about a population based on sample results.**

A. True

B. False

ANSWER: A

6. **Qualitative variable is a variable that cannot assume numerical values but is classified into two or more categories.**
   A. True
   B. False

ANSWER: A

7. **Quantitative variable is a variable that can be measured numerically.**
   A. True
   B. False

ANSWER: A

8. **Population is a sample drawn in such a way that each element of the population has some chance of being included in the sample.**
   A. True
   B. False

ANSWER: B

9. Representative sample is a sample that contains the same characteristics as the corresponding population.

A. True

B. False

ANSWER: A

10. Sample is a portion of the population of interest.

A. True

B. False

ANSWER: A

11. Statistics is a group of methods used to collect, analyze, present, and interpret data and to make decisions.

A. True

B. False

ANSWER: A

12. Variable is a characteristic under study or investigation that assumes different values for different elements.

A. True

B. False

ANSWER: A

13. ………… Can only assume certain values and there are usually" gaps" between values

A. Data

B. Discrete variable

C. Continuous variable

D. value

ANSWER: B

14. What will be the value of $\boldsymbol{P(not\,E)}$ if $\boldsymbol{P(E) = 0.07}$?

A. 0.90

B. 0.007

C. 0.93

D. 10.72

ANSWER: C

15. What will be the probability of getting odd numbers if a die is thrown?

A. 1/2

B. 2

C. 4/2

D. 5/2

ANSWER: A

16. If you roll two dice, the probability of getting a sum as 11 is

A. 2/18

B. 1/18

C. 4

D. 1/36

ANSWER: B

17. The probability of getting two tails when two coins are tossed is -

A. 1/6

B. 1/2

C. 1/3

D. 1/4

ANSWER: D

18. What is the probability of getting the sum as a prime number if two dice are thrown?

A. 5/24

B. 5/12

C. 5/30

D. 1/4

ANSWER: B

19. What is the probability of getting at least one head if three unbiased coins are tossed?

a. 7/8

b. 1/2

c. 5/8

d. 8/9

ANSWER: A

20. A card is drawn from a pack of 52 cards. What is the probability of getting a king of a black suit?

A. 1/26

B. 1/52

C. 3/26

D. 7/52

ANSWER: A

21. What is the probability of drawing an ace from a pack of 52 cards?

A. 4/13

B. 1/13

C. 1/52

D. None of the above

Answer: **B**

22. Which of the following probability cannot exist?

A. 2/5

B. -1.5

C. 0.7

D. None of the above

ANSWER: B

23. What will be the probability of an impossible event?

A. 0
B. 1
C. Infinity
D. None of the above

ANSWER: A

24. If three coins are tossed simultaneously, what is the probability of getting two heads together?

A. 3/8

B. 1/8

C. 5/8

D. None of the above

ANSWER: A

25. A card is drawn from a pack of 52 cards. What is the probability that it is a face card (King, Queen, and Jack only)?

A. 1/26

B. 2/13

C. 1/13

D. 3/13

ANSWER: D

26. If P is the probability of an event, what is the probability of its complement?

A. 1 - 1/P

B. P - 1

C. 1 - P

D. None of the above

ANSWER: C

27. In a Binomial Distribution, if 'n' is the number of trials and 'p' is the probability of success, then the mean value is given by ___________

A. np

B. n

C. p

D. np(1-p)

Answer: A

28. In a Binomial Distribution, if p, q and n are probability of success, failure and number of trials respectively then variance is given by

a) np

b) npq

c) np^2q

d) npq^2

Answer: b

29. If 'X' is a random variable, taking values 'x', probability of success and failure being 'p' and 'q' respectively and 'n' trials being conducted, then what is the probability that 'X' takes values 'x'? Use Binomial Distribution

A. $P(X = x) = {}^{n}C_{x}\ p^{x}\ q^{x}$

B. $P(X = x) = {}^{n}C_{x}\ p^{x}\ q^{(n-x)}$

C. $P(X = x) = {}^{x}C_{n}\ q^{x}\ p^{(n-x)}$

D. $P(x = x) = {}^{x}C_{n}\ p^{n}\ q^{x}$

Answer: B

30. Mutually Exclusive events ___________

A. Contain all sample points

B. Contain all common sample points

C. Does not contain any sample point

D. Does not contain any common sample point

Answer: D

31. The expected value of a discrete random variable 'x' is given by ________

A. P(x)

B. ∑ P(x)

C. ∑ x P(x)

D. 1

Answer: C

32. If 'X' is a continuous random variable, then the expected value is given by

A. P(X)

B. B. $\sum$ x P(x)

C. $\int$ X P(X)

D. D. No value such as expected value

Answer: C

33. If the probability that a bomb dropped from a place will strike the target is 60% and if 10 bombs are dropped, find mean and variance?

A. 0.6, 0.24

B. 6, 2.4

C. 0.4, 0.16

D. 4, 1.6

Answer: B

34. What is the mean and variance for standard normal distribution?

A. Mean is 0 and variance is 1

B. Mean is 1 and variance is 0

C. Mean is 0 and variance is ∞

D. Mean is ∞ and variance is 0

Answer: A

35. The binomial distribution depends on which of the following?

A. Mean and standard deviation
B. Sample size and probability of success
C. Standard deviation and number of successes
D. Mean and probability of success

Answer: B

36. The normal distribution depends on which of the following?

A. Mean and standard deviation
B. Sample size and probability of success
C. Standard deviation and number of successes
D. Mean and probability of success

Answer: A

37. Normal Distribution is applied for

A. Continuous Random Distribution

B. Discrete Random Variable

C. Irregular Random Variable

D. Uncertain Random Variable

Answer: A

38. The shape of the Normal Curve is

A. Bell Shaped

B. Flat

C. Circular

D. Spiked

Answer: A

Question (4) Choose the correct answer

1. A statement made about a population for testing purpose is called?

A. Statistic

B. Hypothesis

C. Level of Significance

D. Test-Statistic

ANSWER B

2. **If the assumed hypothesis is tested for rejection considering it to be true is called?**
   A. Null Hypothesis
   B. Statistical Hypothesis
   C. Simple Hypothesis
   D. Composite Hypothesis

ANSWER A

3. **The point where the Null Hypothesis gets rejected is called as?**
   A. Significant Value
   B. Rejection Value
   C. Acceptance Value
   D. Critical Value

ANSWER D

4. **Type 1 error occurs when?**
   A. We reject H_0 if it is True
   B. We reject H_0 if it is False
   C. We accept H_0 if it is True
   D. We accept H_0 if it is False

ANSWER A

5. A sample size is considered large in which of the following cases?

A. n > or = 30

B. n > or = 50

C. n < or = 30

D. n < or = 50

ANSWER A

6. The population distribution is the technique of collecting information from a portion of the population.

A. True

B. False

ANSWER B

7. A simple random sample is a sample that is selected in such a way that each member of the population has the same being included in the sample.

A. True

B. False

ANSWER A

8. **The probability distribution of a sample statistics is called its sampling distribution.**
   A. True
   B. False

ANSWER A

9. **The assignment of value(s) to a population parameter based on a value of the corresponding sample statistics is sample survey.**
   A. True
   B. False

ANSWER B

10. **The value(s) assigned to a population parameter based on the value of a sample statistic is called an estimator.**
    A. True
    B. False

ANSWER B

11. **The assignment of value(s) to a population parameter based on a value of the corresponding sample statistics is called estimation.**

A. True

B. False

ANSWER A

12. The value of a sample statistic that is used to estimate a population parameter is called a point estimate.

A. True

B. False

ANSWER A

Question (4)

In each of 4 races, the Democrats have a 60% chance of winning. Assuming that the races are independent of each other, what is the probability that the Democrats will win at least 1 race? (5 points)

- $n = 4, \quad p = 0.60, \quad 1 - p = 0.40$

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

$$= 1 - \left[\frac{4!}{0!\,4!} 0.60^0 0.40^4\right] = 0.9744$$

Let x be a random variable that is normally distributed with a mean of 25 and a standard deviation of 4. Find $P(18 < x < 32)$?

$$\mu = 25, \quad \sigma = 4$$

$$P\left(\frac{18-25}{4} < \frac{x-\mu}{\sigma} < \frac{32-25}{4}\right)$$

$$= P(-1.75 < Z < 1.75)$$

$$= 0.4599 + 0.4599 = 0.9198$$

Question (5)

An English professor counted the number of misspelled words on an essay he recently assigned, for his class of 40 students; the mean number of misspelled words was 6.05 and the sample standard deviation 2.44.

- What is the estimated mean number of misspelled words in the population? What is that estimate called?

- Construct 95% confidence interval for the mean number of misspelled words in the population of students.
- Interpret your findings.

Solution

- $\bar{X} = 6.05,\ Point\ estimate$

$$n = 40,\ \bar{X} = 6.05,\ \ S = 2.44,\ \ Z = 1.96$$

- $\bar{X} \mp ZS_{\bar{X}} = 6.05 \mp \left[1.96 \times \frac{2.44}{\sqrt{40}}\right] = 6.05 \mp \left[1.96 \times \frac{2.44}{6.324}\right]$

$6.05 - [1.96 \times 0.3858] = 5.294 \qquad L.L$

$6.05 + [1.96 \times 0.3858] = 6.806 \qquad U.L$

$$5.294 \leq \mu \leq 6.806$$

- The population mean falls between 5.294 as a lower limit, and 6.806 as a higher limit with confidence level 95%.

References

1. Mann, S. Statistics for business and economics. Wiley & Sons, 1995.
2. Mendenhall, W., Beaver, J., and Beaver, M. introduction to probability and statistics, 13^{th} ed. Cengage Learning, 2009.
3. Boewman, L., O'Connell, R,. Applied Statistics. Irwin, 1997.